Generalization Analysis for Contrastive Representation Learning

Yunwen Lei [1] **Tianbao Yang** [2] **Yiming Ying** [3] **Ding-Xuan Zhou** [4]

Abstract

Recently, contrastive learning has found impressive success in advancing the state of the art in solving various machine learning tasks. However, the existing generalization analysis is very limited or even not meaningful. In particular, the existing generalization error bounds depend linearly on the number k of negative examples while it was widely shown in practice that choosing a large k is necessary to guarantee good generalization of contrastive learning in downstream tasks. In this paper, we establish novel generalization bounds for contrastive learning which do not depend on k, up to logarithmic terms. Our analysis uses structural results on empirical covering numbers and Rademacher complexities to exploit the Lipschitz continuity of loss functions. For self-bounding Lipschitz loss functions, we further improve our results by developing optimistic bounds which imply fast rates in a low noise condition. We apply our results to learning with both linear representation and nonlinear representation by deep neural networks, for both of which we derive Rademacher complexity bounds to get improved generalization bounds.

1. Introduction

The performance of machine learning (ML) models often depends largely on the representation of data, which motivates a resurgence of contrastive representation learning (CRL) to learn a representation function $f : \mathcal{X} \mapsto \mathbb{R}^d$ from unsupervised data (Chen et al., 2020; Khosla et al., 2020; He et al., 2020). The basic idea is to pull together similar pairs $(\mathbf{x}, \mathbf{x}^+)$ and push apart disimilar pairs $(\mathbf{x}, \mathbf{x}^-)$ in an embedding space, which can be formulated as minimizing the following objective (Chen et al., 2020; Oord et al., 2018)

$$\mathbb{E}_{\mathbf{x},\mathbf{x}^+,\{\mathbf{x}_i^-\}_{i=1}^k} \log\Big(1+\sum_{i=1}^{k} \exp\left(-f(\mathbf{x})^\top\left(f(\mathbf{x}^+)-f(\mathbf{x}_i^-)\right)\right)\Big),$$

where k is the number of negative examples. The hope is that the learned representation $f(\mathbf{x})$ would capture the latent structure and be beneficial to other downstream learning tasks (Arora et al., 2019; Tosh et al., 2021a). CRL has achieved impressive empirical performance in advancing the state-of-the-art performance in various domains such as computer vision (He et al., 2020; Caron et al., 2020; Chen et al., 2020; Caron et al., 2020) and natural language processing (Brown et al., 2020; Gao et al., 2021; Radford et al., 2021).

The empirical success of CRL motivates a natural question on theoretically understanding how the learned representation adapts to the downstream tasks, i.e.,

How would the generalization behavior of downstream ML models benefit from the representation function built from positive and negative pairs? Especially, how would the number of negative examples affect the learning performance?

Arora et al. (2019) provided an attempt to answer the above questions by developing a theoretical framework to study CRL. They first gave generalization bounds for a learned representation function in terms of Rademacher complexities. Then, they showed that this generalization behavior measured by an unsupervised loss guarantees the generalization behavior of a linear classifier in the downstream classification task. However, the generalization bounds there enjoy a linear dependency on k, which would not be effective if k is large. Moreover, this is not consistent with many studies which show a large number of negative examples (Chen et al., 2020; Tian et al., 2020a; Hénaff et al., 2020; Khosla et al., 2020) is necessary for good generalization performance. For example, the work (He et al., 2020) used 65536 negative examples in unsupervised visual representation learning, for which the existing analysis requires $n \geq (65536)^2 d$ training examples to get non-vacuous bounds (Arora et al., 2019). Yuan et al. (2022) has demonstrated the benefits using all negative data for each anchor data for CRL and proposed an efficient algorithm for optimizing global contrastive loss. Therefore, the existing

[1]Department of Mathematics, The University of Hong Kong [2]Department of Computer Science and Engineering, Texas A&M University [3]Department of Mathematics and Statistics, State University of New York at Albany [4]School of Mathematics and Statistics, University of Sydney. Correspondence to: Yiming Ying <yying@albany.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Assumption	Arora et al.'19	Ours
	$O\left(\frac{k\sqrt{d}}{\sqrt{n}}\right)$	$\widetilde{O}\left(\frac{\sqrt{d}}{\sqrt{n}}\right)$
low noise	$O\left(\frac{k\sqrt{d}}{\sqrt{n}}\right)$	$\widetilde{O}\left(\frac{d}{n}\right)$

Table 1. Comparison between our generalization bounds and those in Arora et al. (2019) for the logistic loss. Here d is the number of learned features. The notation $\widetilde{O}$ ignores log factors.

analysis does not fully answer the question on the super performance of CRL to downstream tasks as already shown in many applications.

In this paper, we aim to further deepen our understanding of CRL by fully exploiting the Lipschitz continuity of loss functions. Our contributions are listed as follows.

1. We develop generalization error bounds for CRL. We consider three types of loss functions: ℓ_2-Lipschitz loss, ℓ_∞-Lipschitz loss and self-bounding Lipschitz loss. For ℓ_2-Lipschitz loss, we develop a generalization bound with a square-root dependency on k by two applications of vector-contraction lemmas on Rademacher complexities, which improves the existing bound by a factor of $\sqrt{k}$ (Arora et al., 2019). For ℓ_∞-Lipschitz loss, we develop generalization bounds which does not depend on k, up to some logarithmic terms, by approximating the arguments of loss functions via expanding the original dataset by a factor of k to fully exploit the Lipschitz continuity. For self-bounding Lipschitz loss, we develop optimistic bounds involving the training errors, which can imply fast rates under a low noise setting. All of our generalization bounds involve Rademacher complexities of feature classes, which preserve the coupling among different features.

2. We then apply our general result to two unsupervised representation learning problems: learning with linear features and learning with nonlinear features via deep neural networks (DNNs). For learning with linear features, we consider two regularization schemes, i.e., p-norm regularizer and Schatten-norm regularizer. For learning with nonlinear features, we develop Rademacher complexity and generalization bounds with a square-root dependency on the depth of DNNs. To this aim, we adapt the technique in Golowich et al. (2018) by using a different moment generalization function to capture the coupling among different features.

3. Finally, we apply our results on representation learning to the generalization analysis of downstream classification problems, which outperforms the existing results by a factor of k (ignoring a log factor).

The remaining parts of the paper are organized as follows. Section 2 reviews the related work, and Section 3 provides the problem formulation. We give generalization bounds for CRL in Section 4 for three types of loss functions, which are then applied to learning with both linear and nonlinear features in Section 5. Conclusions are given in Section 6.

2. Related Work

The most related work is the generalization analysis of CRL in Arora et al. (2019), where the authors developed generalization bounds for unsupervised errors in terms of Rademacher complexity of representation function classes. Based on this, they further studied the performance of linear classifiers on the learned features. In particular, they considered the mean classifier where the weight for a class label is the mean of the representation of corresponding inputs. A major result in Arora et al. (2019) is to show that the classification errors of the mean classifier can be bounded by the unsupervised errors of learned representation functions. This shows that the downstream classification task can benefit from a learned representation function with a low unsupervised error.

The above work motivates several interesting theoretical study of CRL. Nozawa et al. (2020) studied CRL in a PAC-Bayesian setting, which aims to learn a posterior distribution of representation functions. Nozawa et al. (2020) derived PAC-Bayesian bounds for the posterior distribution and applied it to get PAC-Bayesian bounds for the mean-classifier, which relaxes the i.i.d. assumption. Negative examples in the framework (Arora et al., 2019) are typically taken to be randomly sampled datapoints, which may actually have the same label of the point of interest. This introduces a bias in the objective function of CRL, which leads to performance drops in practice. Motivated by this, Chuang et al. (2020) introduced a debiased CRL algorithm by building an approximation of unbiased error in CRL, and developed generalization guarantees for the downstream classification. Ash et al. (2022) refines the connection in Arora et al. (2019) by removing the collision probability in the denominator, which motivates the discussion on selecting the optimal number of negative examples. Nozawa & Sato (2021) improves the analysis Arora et al. (2019) by giving a generic transfer theorem between unsupervised loss and supervised loss, which exhibits a coverage-collision trade-off due to the number of negative examples.

Several researchers studied CRL from other perspectives. Lee et al. (2021) proposed to learn a representation function f to minimize $\mathbb{E}_{(X_1,X_2)}[\|X_2 - f(X_1)\|_2^2]$, where X_1, X_2 are unlabeled input and pretext target. Under an approximate conditional independency assumption, the authors showed that a linear function based on the learned representation approximates the true predictor on downstream problems. In a generative modeling setup, Tosh et al. (2021b) proposed to learn representation functions by a landmark embedding procedure, which can reveal the underlying topic posterior information. Tosh et al. (2021a) studied CRL in a multi-view setting with two views available for each datum. Under

an assumption on the redundancy between the two views, the authors showed that low-dimensional representation can achieve near optimal downstream performance with linear models. HaoChen et al. (2021) studied self-supervised learning from the perspective of spectral clustering based on a population augmentation graph, and proposed a spectral contrastive loss. They further developed generalization bounds for both representation learning and the downstream classification. This result is improved in a recent work by developing a guarantee to incorporate the representation function class (Saunshi et al., 2022). Wang et al. (2022) removes the assumption on the conditional independency, and proposes a guarantee on the downstream performance from the viewpoint of augmentation overlap. Bao et al. (2022) shows that the contrastive loss can be viewed as a surrogate objective of the downstream loss by building upper and lower bounds for downstream classification errors. There are also recent work on theoretical analysis of representation learning via gradient-descent dynamics (Lee et al., 2021; Tian et al., 2020b), mutual information (Tsai et al., 2020), alignment of representations (Wang & Isola, 2020), and causality (Mitrovic et al., 2020). CRL is related to metric learning, for which generalization bounds have been studied in the literature (Cao et al., 2016).

3. Problem Formulation

Let $\mathcal{X}$ denote the space of all possible datapoints. In CRL, we are given several *similar* data in the form of pairs $(\mathbf{x}, \mathbf{x}^+)$ drawn from a distribution $\mathcal{D}_{sim}$ and *negative* data $\mathbf{x}_1^-, \mathbf{x}_2^-, \ldots, \mathbf{x}_k^-$ drawn from a distribution $\mathcal{D}_{neg}$ unrelated to $\mathbf{x}$. Our aim is to learn a feature map $f : \mathcal{X} \mapsto \mathbb{R}^d$ from a class of representation functions $\mathcal{F} = \{f : \|f(\cdot)\|_2 \leq R\}$ for some $R > 0$, where $\|\cdot\|_2$ denotes the Euclidean norm. Here $d \in \mathbb{N}$ denotes the number of features.

We follow the framework in Arora et al. (2019) to define the distribution $\mathcal{D}_{sim}$ and the distribution $\mathcal{D}_{neg}$. Let $\mathcal{C}$ denote the set of all latent classes and for each class $c \in \mathcal{C}$ we assume there is a probability distribution $\mathcal{D}_c$ over $\mathcal{X}$, which quantifies the relevance of $\mathbf{x}$ to the class c. We assume there is a probability distribution ρ defined over $\mathcal{C}$. Then we define $\mathcal{D}_{sim}(\mathbf{x}, \mathbf{x}^+)$ and $\mathcal{D}_{neg}(\mathbf{x}^-)$ as follows

$$\mathcal{D}_{sim}(\mathbf{x}, \mathbf{x}^+) = \mathbb{E}_{c\sim\rho}\big[\mathcal{D}_c(\mathbf{x})\mathcal{D}_c(\mathbf{x}^+)\big],$$
$$\mathcal{D}_{neg}(\mathbf{x}^-) = \mathbb{E}_{c\sim\rho}\big[\mathcal{D}_c(\mathbf{x}^-)\big].$$

Intuitively, $\mathcal{D}_{sim}(\mathbf{x}, \mathbf{x}^+)$ measures the probability of $\mathbf{x}$ and $\mathbf{x}^+$ being drawn from the same class $c \sim \rho$, while $\mathcal{D}_{neg}(\mathbf{x}^-)$ measures the probability of drawing an un-relevant $\mathbf{x}^-$. Let $(\mathbf{x}_j, \mathbf{x}_j^+) \sim \mathcal{D}_{sim}$ and $(\mathbf{x}_{j1}^-, \ldots, \mathbf{x}_{jk}^-) \sim \mathcal{D}_{neg}, j \in [n] := \{1, \ldots, n\}$, where k denotes the number of negative examples. We collect these training examples into a dataset

$$S = \Big\{(\mathbf{x}_1, \mathbf{x}_1^+, \mathbf{x}_{11}^-, \ldots, \mathbf{x}_{1k}^-), (\mathbf{x}_2, \mathbf{x}_2^+, \mathbf{x}_{21}^-, \ldots, \mathbf{x}_{2k}^-), \ldots, (\mathbf{x}_n, \mathbf{x}_n^+, \mathbf{x}_{n1}^-, \ldots, \mathbf{x}_{nk}^-)\Big\}. \quad (3.1)$$

Given a representation function f, we can measure its performance by building a classifier based on this representation and computing the accuracy of the classifier. To this aim, we define a $(K+1)$-way supervised task $\mathcal{T}$ consisting of distinct classes $\{c_1, \ldots, c_{K+1}\} \subseteq \mathcal{C}$. The examples for this supervised task are drawn by the following process:

We first draw a label $c \in \mathcal{T} = \{c_1, \ldots, c_{K+1}\}$ from a distribution $\mathcal{D}_{\mathcal{T}}$ over $\mathcal{T}$, after which we draw an example $\mathbf{x}$ from $\mathcal{D}_c$. This defines the following distribution over labeled pairs $(\mathbf{x}, c)$: $\mathcal{D}_{\mathcal{T}}(\mathbf{x}, c) = \mathcal{D}_c(\mathbf{x})\mathcal{D}_{\mathcal{T}}(c)$. Since there is a label for each example, we can build a multi-class classifier $g : \mathcal{X} \mapsto \mathbb{R}^{K+1}$ for $\mathcal{T}$, where $g_c(\mathbf{x})$ measures the "likelihood" of assigning the class label c to the example $\mathbf{x}$. The loss of g on a point $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{T}$ can be measured by $\ell_s\big(\{g(\mathbf{x})_y - g(\mathbf{x})_{y'}\}_{y'\neq y}\big)$, where $\ell_s : \mathbb{R}^K \mapsto \mathbb{R}^+$. We quantify the performance of a classifier g on the task $\mathcal{T}$ by the supervised loss. By minimizing the supervised loss, we want to build a classifier whose component associated to the correct label is largest.

Definition 3.1 (Supervised loss)**.** Let $g : \mathcal{X} \mapsto \mathbb{R}^{K+1}$ be a multi-class classifier. The supervised loss of g is defined as

$$L_{sup}(\mathcal{T}, g) := \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{D}_{\mathcal{T}}}\Big[\ell_s\big(\{g(\mathbf{x})_c - g(\mathbf{x})_{c'}\}_{c'\neq c}\big)\Big].$$

For CRL, we often consider g as a linear classifier based on the learned representation f, i.e., $g(\mathbf{x}) = Wf(\mathbf{x})$, where $W \in \mathbb{R}^{(K+1)\times d}$. Then the performance of the representation function $f(\mathbf{x})$ can be quantified by the accuracy of the best linear classifier on the representation $f(\mathbf{x})$:

$$L_{sup}(\mathcal{T}, f) = \min_{W\in\mathbb{R}^{(K+1)\times d}} L_{sup}(\mathcal{T}, Wf).$$

To find a good representation f based on unsupervised dataset S, we need to introduce the concept of unsupervised loss functions. Let $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ be a loss function for which popular choices include the hinge loss

$$\ell(\mathbf{v}) = \max\big\{0, 1 + \max_{i\in[k]}\{-v_i\}\big\} \quad (3.2)$$

and the logistic loss

$$\ell(\mathbf{v}) = \log\big(1 + \sum_{i\in[k]} \exp(-v_i)\big). \quad (3.3)$$

Let $f(\mathbf{x})^\top$ denote the transpose of $f(\mathbf{x})$.

Definition 3.2 (Unsupervised error)**.** The population unsupervised error is defined as

$$L_{un}(f) := \mathbb{E}\big[\ell\big(\big\{f(\mathbf{x})^\top(f(\mathbf{x}^+) - f(\mathbf{x}_i^-))\big\}_{i=1}^k\big)\big].$$

The empirical unsupervised error with S is defined as

$$\hat{L}_{un}(f) := \frac{1}{n}\sum_{j=1}^{n} \ell\big(\{f(\mathbf{x}_j)^\top (f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-))\}_{i=1}^k\big).$$

A natural algorithm is to find among $\mathcal{F}$ the function with the minimal empirical unsupervised loss, i.e., $\hat{f} := \arg\min_{f\in\mathcal{F}} \hat{L}_{un}(f)$. This function can then be used for the downstream supervised learning task, e.g., to find a linear classifier $g(\mathbf{x}) = Wf(\mathbf{x})$ indexed by $W \in \mathbb{R}^{(K+1)\times d}$.

4. Generalization Error Bounds

In this paper, we are interested in the performance of $\hat{f}$ on testing, i.e., how the empirical behavior of $\hat{f}$ on S would generalize well to testing examples. Specifically, we will control $L_{un}(\hat{f}) - \hat{L}_{un}(\hat{f})$. Since $\hat{f}$ depends on the dataset S, we need to control the uniform deviation between population unsupervised error and empirical unsupervised error over the function class $\mathcal{F}$, which depends on the complexity of $\mathcal{F}$. In this paper, we will use Rademacher complexity to quantify the complexity of $\mathcal{F}$ (Bartlett & Mendelson, 2002).

Definition 4.1 (Rademacher Complexity)**.** Let $\widetilde{\mathcal{F}}$ be a class of real-valued functions over a space $\mathcal{Z}$ and $\widetilde{S} = \{\mathbf{z}_i\}_{i=1}^n \subseteq \mathcal{Z}$. The *empirical* Rademacher complexity of $\widetilde{\mathcal{F}}$ with respect to (w.r.t.) $\widetilde{S}$ is defined as $\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}}) = \mathbb{E}_{\boldsymbol{\epsilon}}\big[\sup_{f\in\widetilde{\mathcal{F}}} \frac{1}{n}\sum_{i\in[n]} \epsilon_i f(\mathbf{z}_i)\big]$, where $\boldsymbol{\epsilon} = (\epsilon_i)_{i\in[n]} \sim \{\pm 1\}^n$ are independent Rademacher variables. We define the *worst-case* Rademacher complexity as $\mathfrak{R}_{\mathcal{Z},n}(\widetilde{\mathcal{F}}) = \sup_{\widetilde{S}\subseteq\mathcal{Z}:|\widetilde{S}|=n} \mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}})$, where $|\widetilde{S}|$ is the cardinality of $\widetilde{S}$.

For any $f \in \mathcal{F}$, we introduce $g_f : \mathcal{X}^{k+2} \mapsto \mathbb{R}$ as follows

$$g_f(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \ldots, \mathbf{x}_k^-) = \ell\big(\{f(\mathbf{x})^\top \big(f(\mathbf{x}^+) - f(\mathbf{x}_i^-)\big)\}_{i=1}^k\big).$$

It is then clear that

$$L_{un}(f) - \hat{L}_{un}(f) = \mathbb{E}_{\mathbf{x},\mathbf{x}^+,\mathbf{x}_1^-,\ldots,\mathbf{x}_k^-}\big[g_f(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \ldots, \mathbf{x}_k^-)\big] - \frac{1}{n}\sum_{j\in[n]} g_f(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_{j1}^-, \ldots, \mathbf{x}_{jk}^-).$$

Results in learning theory show that we can bound $L_{un}(\hat{f}) - \hat{L}_{un}(\hat{f})$ by $\mathfrak{R}_S(\mathcal{G})$ (Bartlett & Mendelson, 2002), where

$$\mathcal{G} = \big\{(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \ldots, \mathbf{x}_k^-) \mapsto g_f(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \ldots, \mathbf{x}_k^-) : f \in \mathcal{F}\big\}.$$

Note functions in $\mathcal{G}$ involve the nonlinear function $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$, which introduces difficulties in the complexity analysis. Our key idea is to use the Lipschitz continuity of ℓ to reduce the complexity of $\mathcal{G}$ to the complexity of another function class without ℓ. Since the arguments in ℓ are vectors, we can have different definition of Lipschitz continuity w.r.t. different norms (Lei et al., 2015; Tewari & Chaudhuri, 2015; Lei et al., 2019; Foster & Rakhlin, 2019; Mustafa et al., 2022). For any $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^k$ and $p \geq 1$, we define the ℓ_p-norm as $\|\mathbf{a}\|_p = \big(\sum_{i=1}^n |a_i|^p\big)^{\frac{1}{p}}$.

Definition 4.2 (Lipschitz continuity)**.** We say $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ is G-Lipschitz w.r.t. the ℓ_p-norm iff

$$|\ell(\mathbf{a}) - \ell(\mathbf{a}')| \leq G\|\mathbf{a} - \mathbf{a}'\|_p, \quad \forall \mathbf{a}, \mathbf{a}' \in \mathbb{R}^k.$$

In this paper, we are particularly interested in the Lipschitz continuity w.r.t. either the ℓ_2-norm or the ℓ_∞-norm. According to Proposition G.1, the loss functions defined in Eq. (3.2) and Eq. (3.3) are 1-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$, and 1-Lipschitz continuous w.r.t. $\|\cdot\|_2$ (Lei et al., 2019).

Note each component of the arguments in ℓ are of the form $f(\mathbf{x})^\top(f(\mathbf{x}^+) - f(\mathbf{x}^-))$. This motivates the definition of the following function class

$$\mathcal{H} = \Big\{h_f(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = f(\mathbf{x})^\top\big(f(\mathbf{x}^+) - f(\mathbf{x}^-)\big) : f \in \mathcal{F}\Big\}.$$

As we will see in the analysis, the complexity of $\mathcal{G}$ is closely related to that of $\mathcal{H}$. Therefore, we first show how to control the complexity of $\mathcal{H}$. In the following lemma, we provide Rademacher complexity bounds of $\mathcal{H}$ w.r.t. a general dataset S' of cardinality n. We will use a vector-contraction lemma to prove it (Maurer, 2016). The basic idea is to notice the Lipschitz continuity of the map $(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) \mapsto \mathbf{x}^\top(\mathbf{x}^+ - \mathbf{x}^-)$ w.r.t. $\|\cdot\|_2$ on $\mathcal{X}^3$. The proof is given in Section B.

Lemma 4.3. *Let $n \in \mathbb{N}$ and $S' = \{(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_j^-) : j \in [n]\}$. Assume $\|f(\mathbf{x})\|_2 \leq R$ for any $f \in \mathcal{F}$ and $\mathbf{x} \in S'$. Then*

$$\mathfrak{R}_{S'}(\mathcal{H}) \leq \frac{\sqrt{12}R}{n}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n\times\{\pm1\}^d\times\{\pm1\}^3}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\Big(\epsilon_{j,t,1}f_t(\mathbf{x}_j) + \epsilon_{j,t,2}f_t(\mathbf{x}_j^+) + \epsilon_{j,t,3}f_t(\mathbf{x}_j^-)\Big)\Big],$$

where $f_t(\mathbf{x})$ is the t-th component of $f(\mathbf{x}) \in \mathbb{R}^d$.

Remark 4.4. We compare Lemma 4.3 with the following Rademacher complexity bound in HaoChen et al. (2021)

$$\mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{n}}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j f(\mathbf{x}_j)^\top f(\mathbf{x}_j^+)\Big] \leq d\max_{t\in[d]}\mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{n}}\Big[\sup_{f_t\in\mathcal{F}_t}\sum_{j\in[n]}\epsilon_j f_t(\mathbf{x}_j)f_t(\mathbf{x}_j^+)\Big], \quad (4.1)$$

where $\mathcal{F}_t = \big\{\mathbf{x} \mapsto f_t(\mathbf{x}) : f \in \mathcal{F}\big\}$. As a comparison, our analysis in Lemma 4.3 can imply the following bound

$$\mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{n}}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j f(\mathbf{x}_j)^\top f(\mathbf{x}_j^+)\Big] \leq 2R\mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{2nd}}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\mathbf{x}_j) + \epsilon_{j,t,2}f_t(\mathbf{x}_j^+)\big)\Big].$$

Eq. (4.1) decouples the relationship among different features since the maximization over $t \in [d]$ is outside of the

expectation operator. As a comparison, our result preserves this coupling since the summation over $t \in [d]$ is inside the supermum over $f \in \mathcal{F}$. This preservation of coupling has an effect on the bound. Indeed, it is expected that

$$\mathbb{E}_{\epsilon\sim\{\pm 1\}^{2nd}}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\mathbf{x}_j)+\epsilon_{j,t,2}f_t(\mathbf{x}_j^+)\big)\Big] = O\Big(\sqrt{d}\mathbb{E}_{\epsilon\sim\{\pm 1\}^{2n}}\Big[\sup_{f_t\in\mathcal{F}_t}\sum_{j\in[n]}\big(\epsilon_{j,1}f_t(\mathbf{x}_j)+\epsilon_{j,2}f_t(\mathbf{x}_j^+)\big)\Big]\Big).$$

In this case, our result implies a bound with a better dependency on d as compared to Eq. (4.1) (the factor of d in Eq. (4.1) is replaced by $\sqrt{d}$ here). We can plug our bound into the analysis in HaoChen et al. (2021) to improve their results. In Section A we will give a specific example where our bound can outperform Eq. (4.1) by a factor of $\sqrt{d}$.

Remark 4.5. Lemma 4.3 requires an assumption $\|f(\mathbf{x})\|_2 \leq R$. This assumption can be achieved by adding a projection operator as $f(\mathbf{x}) = \mathcal{P}_R(\tilde{f}(\mathbf{x}))$ for $\tilde{f} \in \mathcal{F}$, where $\mathcal{P}_R$ denotes the projection operator onto the Euclidean ball with radius R around the zero point. According to the inequality $\|\mathcal{P}_R(\tilde{f}(\mathbf{x})) - \mathcal{P}_R(\tilde{f}'(\mathbf{x}))\|_2 \leq \|\tilde{f}(\mathbf{x}) - \tilde{f}'(\mathbf{x})\|_2$, the arguments in the proof indeed show the following inequality with $\mathcal{H} = \big\{h_{\tilde{f}}(\mathbf{x},\mathbf{x}^+,\mathbf{x}^-) = \mathcal{P}_R(\tilde{f}(\mathbf{x}))^\top\big(\mathcal{P}_R(\tilde{f}(\mathbf{x}^+)) - \mathcal{P}_R(\tilde{f}(\mathbf{x}^-))\big) : \tilde{f} \in \mathcal{F}\big\}$:

$$\mathfrak{R}_{S'}(\mathcal{H}) \leq \frac{\sqrt{12}R}{n}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{3nd}}\Big[\sup_{\tilde{f}\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\Big(\epsilon_{j,t,1}\tilde{f}_t(\mathbf{x}_j)+\epsilon_{j,t,2}\tilde{f}_t(\mathbf{x}_j^+)+\epsilon_{j,t,3}\tilde{f}_t(\mathbf{x}_j^-)\Big)\Big].$$

That is, we can add a projection operator over $\mathcal{F}$ to remove the assumption $\|f(\mathbf{x})\|_2 \leq R$.

Loss	Arora et al.'19	Ours
1-ℓ_2-Lipschitz	$\frac{\sqrt{k}\mathfrak{B}}{n}$	$\frac{\mathfrak{A}}{n}$
1-ℓ_∞-Lipschitz	$\frac{\sqrt{k}\mathfrak{B}}{n}$	$\frac{\mathfrak{C}}{n\sqrt{k}}^*$
S.B. 1-Lipschitz	———	$n^{-1}+n^{-2}k^{-1}\mathfrak{C}^2+(n^{-\frac{1}{2}}$ $+n^{-1}k^{-\frac{1}{2}}\mathfrak{C})\hat{L}_{un}^{\frac{1}{2}}(f)^*$

Table 2. Comparison between our generalization bounds and those in Arora et al. (2019). The notation * means we ignore log factors. S.B. means self-bounding. The notations $\mathfrak{A}, \mathfrak{B}$ and $\mathfrak{C}$ are defined in Eq. (4.2), (4.4) and (4.5), which are typically of the same order. Then, our results improve the bounds in Arora et al. (2019) by a factor of $\sqrt{k}$ for ℓ_2-Lipschitz loss, and by a factor of k for ℓ_∞-Lipscthiz loss. For self-bounding loss, we get optimistic bounds.

4.1. ℓ_2 Lipschitz Loss

We first consider the ℓ_2 Lipschitz loss. The following theorem to be proved in Section B gives Rademacher complexity and generalization error bounds for unsupervised loss function classes. We always assume $\ell\big(\big\{f(\mathbf{x})^\top(f(\mathbf{x}^+) - f(\mathbf{x}_i^-))\big\}_{i=1}^k\big) \leq B$ for any $f \in \mathcal{F}$ in this paper.

Theorem 4.6 (Generalization bound: ℓ_2-Lipschitz loss)**.** *Assume $\|f(\mathbf{x})\|_2 \leq R$ for any $f \in \mathcal{F}$ and $\mathbf{x} \in \mathcal{X}$. Let S be defined as in Eq. (3.1). If $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ is G_2-Lipschitz w.r.t. the ℓ_2-norm, then $\mathfrak{R}_S(\mathcal{G}) \leq \frac{\sqrt{24}RG_2\mathfrak{A}}{n}$, where*

$$\mathfrak{A} = \mathbb{E}_{\{\epsilon\}\sim\{\pm1\}^{3nkd}}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{i\in[k]}\sum_{t\in[d]}\Big(\epsilon_{j,i,t,1}f_t(\mathbf{x}_j) + \epsilon_{j,i,t,2}f_t(\mathbf{x}_j^+) + \epsilon_{j,i,t,3}f_t(\mathbf{x}_{ji}^-)\Big)\Big]. \quad (4.2)$$

Furthermore, for any $\delta \in (0,1)$, with probability at least $1-\delta$ the following inequality holds for any $f \in \mathcal{F}$

$$L_{un}(f) - \hat{L}_{un}(f) \leq \frac{4\sqrt{6}RG_2\mathfrak{A}}{n} + 3B\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Remark 4.7. Under the same Lipschitz continuity w.r.t. $\|\cdot\|_2$, the following bound was established in Arora et al. (2019)

$$L_{un}(f) = \hat{L}_{un}(f) + O\Big(\frac{G_2R\sqrt{k}\mathfrak{B}}{n} + B\sqrt{\frac{\log(1/\delta)}{n}}\Big), \quad (4.3)$$

where

$$\mathfrak{B} = \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n\times\{\pm1\}^d\times\{\pm1\}^{k+2}}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\Big(\epsilon_{j,t,k+1}f_t(\mathbf{x}_j) + \epsilon_{j,t,k+2}f_t(\mathbf{x}_j^+) + \sum_{i\in[k]}\epsilon_{j,t,k}f_t(\mathbf{x}_{ji}^-)\Big)\Big]. \quad (4.4)$$

Note $\mathfrak{A}$ and $\mathfrak{B}$ are of the same order. Indeed, the dominating term in the braces of the above equation is $\sum_{i\in[k]}\epsilon_{j,t,k}f_t(\mathbf{x}_{ji}^-)$ and therefore we have

$$\mathfrak{B} \asymp \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\sum_{i\in[k]}\epsilon_{j,t,k}f_t(\mathbf{x}_{ji}^-)\Big].$$

Furthermore, $\mathfrak{A}$ grows also in this order since $\epsilon_{j,i,t,1}f_t(\mathbf{x}_j) + \epsilon_{j,i,t,2}f_t(\mathbf{x}_j^+) + \epsilon_{j,i,t,3}f_t(\mathbf{x}_{ji}^-)$ is of the same order of $\epsilon_{j,i,t,3}f_t(\mathbf{x}_{ji}^-)$. Typically, $\mathfrak{A} \asymp \mathfrak{B} \asymp \sqrt{nkd}$ since there are $O(nkd)$ terms in the summation inside the supremum. In this case, Theorem 4.6 implies a bound $O(\sqrt{kd/n})$, while Eq. (4.3) gives a bound $O(k\sqrt{d}/\sqrt{n})$. It is clear our bound improves the bound in Arora et al. (2019) by a factor of $\sqrt{k}$.

4.2. ℓ_∞ Lipschitz Loss

We now turn to the analysis for the setting with ℓ_∞ Lipschitz continuity assumption, which is more challenging. The following theorem controls the Rademacher complexity of $\mathcal{G}$ w.r.t. the dataset S in terms of the worst-case Rademacher

complexity of $\mathcal{H}$ defined on the set $S_{\mathcal{H}}$, where

$$S_{\mathcal{H}} = \Big\{ \underbrace{(\mathbf{x}_1, \mathbf{x}_1^+, \mathbf{x}_{11}^-), (\mathbf{x}_1, \mathbf{x}_1^+, \mathbf{x}_{12}^-), \ldots, (\mathbf{x}_1, \mathbf{x}_1^+, \mathbf{x}_{1k}^-)}_{\text{induced by the first example}},$$
$$\underbrace{(\mathbf{x}_2, \mathbf{x}_2^+, \mathbf{x}_{21}^-), (\mathbf{x}_2, \mathbf{x}_2^+, \mathbf{x}_{22}^-), \ldots, (\mathbf{x}_2, \mathbf{x}_2^+, \mathbf{x}_{2k}^-)}_{\text{induced by the second example}}, \ldots,$$
$$\underbrace{(\mathbf{x}_n, \mathbf{x}_n^+, \mathbf{x}_{n1}^-), (\mathbf{x}_n, \mathbf{x}_n^+, \mathbf{x}_{n2}^-), \ldots, (\mathbf{x}_n, \mathbf{x}_n^+, \mathbf{x}_{nk}^-)}_{\text{induced by the last example}} \Big\}.$$

As compared to $\mathcal{G}$, the function class $\mathcal{H}$ removes the loss function ℓ and is easier to handle. Our basic idea is to exploit the Lipschitz continuity of ℓ w.r.t. $\|\cdot\|_\infty$: to approximate the function class $\{\ell(v_1(\mathbf{y}), \ldots, v_k(\mathbf{y}))\}$, it suffices to approximate each component $v_j(\mathbf{y}), j \in [k]$. This explains why we expand the set S of cardinality n to the set $S_{\mathcal{H}}$ of cardinality nk. The proof is given in Section C.

Theorem 4.8 (Complexity bound: ℓ_∞-Lipschitz loss)**.** *Assume $\|f(\mathbf{x})\|_2 \leq R$ for any $f \in \mathcal{F}$ and $\mathbf{x} \in \mathcal{X}$. Let S be defined as in Eq. (3.1). If $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ is G-Lipschitz w.r.t. the ℓ_∞-norm, then*

$$\mathfrak{R}_S(\mathcal{G}) \leq 24G(R^2+1)n^{-\frac{1}{2}} + 48G\sqrt{k}\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H}) \times \left(1 + \log(4R^2 n^{\frac{3}{2}}k)\left\lceil \log_2 \frac{R^2\sqrt{n}}{12} \right\rceil\right),$$

where $\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H}) = \max_{\{(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_j^+, \tilde{\mathbf{x}}_j^-)\}_{j\in[nk]} \subseteq S_{\mathcal{H}}} \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nk}}$

$$\Big[\sup_{h\in\mathcal{H}} \frac{1}{nk} \sum_{j\in[nk]} \epsilon_j f(\tilde{\mathbf{x}}_j)^\top (f(\tilde{\mathbf{x}}_j^+) - f(\tilde{\mathbf{x}}_j^-))\Big].$$

Note in $\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H})$ we restrict the domain of functions in $\mathcal{H}$ to $S_{\mathcal{H}}$, and allow an element in $S_{\mathcal{H}}$ to be chosen several times in the above maximization.

We can use Lemma 4.3 to control $\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H})$ in Theorem 4.8, and derive the following generalization error bound. The proof is given in Section C.

Theorem 4.9 (Generalization bound: ℓ_∞-Lipschitz loss)**.** *Let $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ be G-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. Assume $\|f(\mathbf{x})\|_2 \leq R, \delta \in (0,1)$. Then with probability at least $1-\delta$ over S for all $f \in \mathcal{F}$ we have*

$$L_{un}(f) \leq \hat{L}_{un}(f) + 3B\sqrt{\frac{\log(2/\delta)}{2n}} + 48G(R^2+1)n^{-\frac{1}{2}} + \frac{96\sqrt{12}GR}{n\sqrt{k}}\left(1 + \log(4R^2n^{\frac{3}{2}}k)\left\lceil \log_2 \frac{R^2\sqrt{n}}{12}\right\rceil\right)\mathfrak{C},$$

where

$$\mathfrak{C} = \max_{\{(\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_j^+, \tilde{\mathbf{x}}_j^-)\}_{j=1}^{nk} \subseteq S_{\mathcal{H}}} \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nk}\times\{\pm1\}^d\times\{\pm1\}^3} \quad (4.5)$$

$$\Big[\sup_{f\in\mathcal{F}} \sum_{j\in[nk]} \sum_{t\in[d]} \big(\epsilon_{j,t,1} f_t(\tilde{\mathbf{x}}_j) + \epsilon_{j,t,2} f_t(\tilde{\mathbf{x}}_j^+) + \epsilon_{j,t,3} f_t(\tilde{\mathbf{x}}_j^-)\big)\Big].$$

Remark 4.10. We now compare our bound with Eq. (4.3) developed in Arora et al. (2019). It is reasonable to assume $\mathfrak{C}$ and $\mathfrak{B}$ are of the same order. [1] Then, our bound becomes

$$L_{un}(f) = \hat{L}_{un}(f) + O\Big(\frac{GR\mathfrak{B}\log^2(nRk)}{n\sqrt{k}} + B\sqrt{\frac{\log(1/\delta)}{n}}\Big). \quad (4.6)$$

We know if ℓ is G_2-Lipschitz continuous w.r.t. $\|\cdot\|_2$, it is also $\sqrt{k}G_2$-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. Therefore, in the extreme case we have $G = \sqrt{k}G_2$. Even in this extreme case, our bound is of the order $L_{un}(f) = \hat{L}_{un}(f) + O\Big(\frac{G_2 R\mathfrak{B}\log^2(nRk)}{n} + B\sqrt{\frac{\log(1/\delta)}{n}}\Big)$, which improves Eq. (4.3) by a factor of $\sqrt{k}$ up to a logarithmic factor. For popular loss functions defined in Eq. (3.2) and Eq. (3.3), we have $G = G_2 = 1$ and in this case, our bound in Eq. (4.6) improves Eq. (4.3) by a factor of k if we ignore a logarithmic factor.

Remark 4.11. we now give the intuition of our improvements. Arora et al. (2019) considers the 1-Lipschitz continuity of $\ell : \mathbb{R}^k \mapsto \mathbb{R}$ w.r.t. $\|\cdot\|_2$, while we use the 1-Lipschitz continuity of ℓ w.r.t. $\|\cdot\|_\infty$. Note that 1-Lipschitz continuity w.r.t. $\|\cdot\|_2$ is a weaker condition as compared to the 1-Lipschitz continuity w.r.t. $\|\cdot\|_\infty$. Indeed, if ℓ is 1-Lipschitz continuous w.r.t. $\|\cdot\|_2$, then it is also $\sqrt{k}$-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$ with a much larger Lipschitz constant. In our problem, the contrastive loss is 1-Lipschitz continuous w.r.t. both $\|\cdot\|_2$ and $\|\cdot\|_\infty$. Therefore, we can use the stronger assumption on the 1-Lipschitz continuity w.r.t. $\|\cdot\|_\infty$ to save a factor of $\sqrt{k}$. Furthermore, Arora et al. (2019) use the inequality $\|J\|_2 \leq \|J\|_F$, where $J \in \mathbb{R}^{k\times(k+2)d}$, $\|\cdot\|_2$ is the spectral norm and $\|\cdot\|_F$ is the Frobenius norm. The inequality introduces an additional factor of $\sqrt{k}$ since $\|J\|_F$ can be as large as $\sqrt{k}\|J\|_2$. As a comparison, our analysis based on Lipschitz continuity w.r.t. $\|\cdot\|_\infty$ does not introduce any loss in the factor of k (up to a log term), and outperforms Arora et al. (2019) by a factor of k.

4.3. Self-bounding Lipschitz Loss

Finally, we consider a self-bounding Lipschitz condition where the Lipschitz constant depends on the loss function values. This definition was given in Reeve & Kaban (2020).

Definition 4.12 (Self-bounding Lipschitz Continuity)**.** A loss function $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ is said to be G_s-self-bounding Lipschitz continuous w.r.t. ℓ_∞ norm if for any $\boldsymbol{a}, \boldsymbol{a}' \in \mathbb{R}^k$

$$\big|\ell(\boldsymbol{a}) - \ell(\boldsymbol{a}')\big| \leq G_s \max\big\{\ell(\boldsymbol{a}), \ell(\boldsymbol{a}')\big\}^{\frac{1}{2}} \|\boldsymbol{a} - \boldsymbol{a}'\|_\infty.$$

It was shown that the logistic loss given in Eq. (3.3) satisfies the self-bounding Lipschtiz continuity with $G_s = 2$ (Reeve

[1] Indeed, under a typical behavior of Rademacher complexity as $\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \sup_{\boldsymbol{a}\in\mathcal{A}\subset\mathbb{R}^n} \big[\epsilon_i a_i\big] = O(\sqrt{n})$ (Bartlett & Mendelson, 2002), we have $\mathfrak{C} = O(\sqrt{nkd})$ and $\mathfrak{B} = O(\sqrt{nkd})$.

& Kaban, 2020). In the following theorem, we give generalization bounds for learning with self-bounding Lipschitz loss functions. The basic idea is to replace the Lipschitz constant G in Theorem 4.9 with empirical errors by using the self-bounding property. We use $\widetilde{O}$ to hide logarithmic factors. The proof is given in Section D.

Theorem 4.13 (Generalization bound: self-bounding Lipschitz loss)**.** *Let $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ be G_s-self-bounding Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. Assume $\|f(\mathbf{x})\|_2 \leq R, \delta \in (0,1)$. Then with probability at least $1-\delta$ over S we have the following inequality uniformly for all $f \in \mathcal{F}$*

$$L_{un}(f) = \hat{L}_{un}(f) + \widetilde{O}\Big((B + G_s^2R^4)n^{-1} + G_s^2R^2n^{-2}k^{-1}\mathfrak{C}^2\Big)$$
$$+\widetilde{O}\Big((\sqrt{B} + G_sR^2)n^{-\frac{1}{2}} + G_sRn^{-1}k^{-\frac{1}{2}}\mathfrak{C}\Big)\hat{L}_{un}^{\frac{1}{2}}(f)\log^{\frac{1}{2}}(1/\delta).$$

Remark 4.14. Theorem 4.13 gives optimistic generalization bounds in the sense that the upper bounds depend on empirical errors (Srebro et al., 2010). Therefore, the generalization bounds for $L_{un}(f) - \hat{L}_{un}(f)$ would benefit from low training errors. In particular, if $\hat{L}_{un}^{\frac{1}{2}}(f) = 0$, Theorem 4.13 implies generalization bounds

$$L_{un}(f) = \hat{L}_{un}(f) + \widetilde{O}\Big(Bn^{-1} + G_s^2R^4n^{-1} + G_s^2R^2n^{-2}k^{-1}\mathfrak{C}^2\Big).$$

Typically, we have $\mathfrak{C} = O(\sqrt{nk})$ and in this case $L_{un}(f) = \hat{L}_{un}(f) + \widetilde{O}\big(Bn^{-1} + G_s^2R^4n^{-1}\big)$. In other words, we get fast-decaying error bounds in an interpolating setting.

5. Applications

To apply Theorem 4.6 and Theorem 4.9, we need to control the term $\mathfrak{A}$ or $\mathfrak{C}$, which is related to the Rademacher complexity of a function class. In this section, we will show how to control $\mathfrak{C}$ for features of the form $\mathbf{x} \mapsto U\mathbf{v}(\mathbf{x})$, where $U \in \mathbb{R}^{d\times d'}$ is a matrix and $\mathbf{v} : \mathcal{X} \mapsto \mathbb{R}^{d'}$. Here $\mathbf{v}$ maps the original data $\mathbf{x} \in \mathcal{X}$ to an intermediate feature in $\mathbb{R}^{d'}$, which is used for all the final features. If $\mathbf{v}$ is the identity map, then we get linear features. If $\mathbf{v}$ is a neural network, then we get nonlinear features. For a norm $\|\cdot\|$ on a matrix, we denote by $\|\cdot\|_*$ its dual norm. The following lemmas to be proved in Section E give general results on Rademacher complexities. Lemma 5.1 gives upper bounds, while Lemma 5.2 gives lower bounds. It is immediate to extend our analysis to control $\mathfrak{A}$. For brevity we ignore such a discussion.

Lemma 5.1 (Upper bound)**.** *Let $d, d' \in \mathbb{N}$. Let $\mathcal{V}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}^{d'}$. Let $\mathcal{F} = \{f(\mathbf{x}) = U\mathbf{v}(\mathbf{x}) : U \in \mathcal{U}, \mathbf{v} \in \mathcal{V}\}$, where $\mathcal{U} = \big\{U = (\mathbf{u}_1, \ldots, \mathbf{u}_d)^\top \in \mathbb{R}^{d\times d'} : \|U^\top\| \leq \Lambda\big\}$ and $f(\mathbf{x}) = U\mathbf{v}(\mathbf{x}) = (\mathbf{u}_1, \ldots, \mathbf{u}_d)^\top\mathbf{v}(\mathbf{x})$. Then*

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]}\sum_{j\in[n]} \epsilon_{j,t} f_t(\mathbf{x}_j) \leq$$
$$\Lambda\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \sup_{\mathbf{v}\in\mathcal{V}} \Big\|\Big(\sum_{j\in[n]}\epsilon_{1,j}\mathbf{v}(\mathbf{x}_j), \ldots, \sum_{j\in[n]}\epsilon_{d,j}\mathbf{v}(\mathbf{x}_j)\Big)\Big\|_*.$$

Lemma 5.2 (Lower bound)**.** *If $\mathcal{F}$ is symmetric in the sense that $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$, then we have*

$$\mathfrak{C} \geq \sqrt{\frac{nk}{2}} \sup_{f\in\mathcal{F}} \max_{(\tilde{\mathbf{x}},\tilde{\mathbf{x}}^+,\tilde{\mathbf{x}}^-)\in S_{\mathcal{H}}} \Big(\|f(\tilde{\mathbf{x}})\|_2^2 + \|f(\tilde{\mathbf{x}}^+)\|_2^2 + \|f(\tilde{\mathbf{x}}^-)\|_2^2\Big)^{\frac{1}{2}}.$$

Note in our definition of $\mathcal{F}$, we ignore the projection operator, i.e., the feature function class should be of the form $f(\mathbf{x}) = \mathcal{P}_R\big(U\mathbf{v}(\mathbf{x})\big)$ to satisfy the assumption $\|f(x)\|_2 \leq R$. According to Remark 4.5, it is easy to extend our analysis here to the case with including the projection operator in the definition of feature function class.

5.1. Linear Features

We first apply Lemma 5.1 to derive Rademacher complexity bounds for learning with linear features. For any $p \geq 1$ and a matrix $W = (\mathbf{w}_1, \ldots, \mathbf{w}_{d'}) \in \mathbb{R}^{d\times d'}$, the $\ell_{2,p}$ norm of W is defined as $\|W\|_{2,p} = \big(\sum_{i\in[d']}\|\mathbf{w}_i\|_2^p\big)^{\frac{1}{p}}$. If $p = 2$, this becomes the Frobenius norm $\|W\|_F$. For any $p \geq 1$, the Schatten-p norm of a matrix $W \in \mathbb{R}^{d\times d'}$ is defined as the ℓ_p-norm of the vector of singular values $(\sigma_1(W), \ldots, \sigma_{\min\{d,d'\}}(W))^\top$ (the singular values are assumed to be sorted in non-increasing order), i.e., $\|W\|_{S_p} := \|\sigma(W)\|_p$. Let p^* be the number satisfying $1/p + 1/p^* = 1$. The following proposition to be proved in Section E.1 gives complexity bounds for learning with linear features.

Proposition 5.3 (Linear representation)**.** *Consider the feature map defined in Lemma 5.1 with $\mathbf{v}(\mathbf{x}) = \mathbf{x}$.*

(a) If $\|\cdot\| = \|\cdot\|_{2,p}$, then $\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]}\sum_{j\in[n]} \epsilon_{t,j} f_t(\mathbf{x}_j)$

$$\leq \min_{q\geq p}\Big\{\Lambda d^{1/q^*}\max(\sqrt{q^*-1}, 1)\Big\}\Big(\sum_{j\in[n]}\|\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}.$$

(b) If $\|\cdot\| = \|\cdot\|_{S_p}$ with $p \leq 2$, then

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]}\sum_{j\in[n]} \epsilon_{t,j} f_t(\mathbf{x}_j) \leq \Lambda 2^{-\frac{1}{4}} \min_{q\in[p,2]} \sqrt{\frac{q^*\pi}{e}} \times$$
$$\max\Big\{\Big\|\Big(d\sum_{j\in[n]}\mathbf{x}_j\mathbf{x}_j^\top\Big)^{\frac{1}{2}}\Big\|_{S_{q^*}}, d^{1/q^*}\Big(\sum_{j\in[n]}\|\mathbf{x}_j\|_2^2\Big)^{1/2}\Big\}.$$

We now plug the above Rademacher complexity bounds into Theorem 4.9 to give generalization error bounds for learning with unsupervised loss. Let $B_x = \max\{\|\mathbf{x}_j\|_2, \|\mathbf{x}_j^+\|_2, \|\mathbf{x}_{jt}^-\|_2 : j \in [n], t \in [k]\}$. Note $\big(\sum_{j\in[nk]}\|\tilde{\mathbf{x}}_j\|_2^2\big)^{\frac{1}{2}} \leq \sqrt{nk}B_x$ for $\tilde{\mathbf{x}}_j$ in the definition of $\mathfrak{C}$, from which and Proposition 5.3 we get the following bound for the case $\mathbf{v}(\mathbf{x}) = \mathbf{x}$ (the definition of $\mathfrak{C}$ involves nk examples, while in Proposition 5.3 we consider n examples):

$$\mathfrak{C} = O\Big(\sqrt{nk}B_x \min_{q\geq p}\big\{\Lambda d^{1/q^*}\max(\sqrt{q^*-1}, 1)\big\}\Big).$$

The following corollary then follows from Theorem 4.9.

Corollary 5.4. *Consider the feature map in Proposition 5.3 with $\|\cdot\| = \|\cdot\|_{2,p}$. Let ℓ be the logistic loss and $\delta \in (0,1)$. Then with probability at least $1-\delta$, for all $f \in \mathcal{F}$*

$$L_{un}(f) - \hat{L}_{un}(f) = \frac{B\log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}} + \widetilde{O}\Big(\frac{GRB_x \min_{q\geq p}\{\Lambda d^{1/q^*}\max(\sqrt{q^*-1},1)\}}{\sqrt{n}}\Big).$$

It is also possible to give generalization bounds for learning with ℓ_2-Lipschitz loss functions, and optimistic generalization bounds for learning with self-bounding Lipschitz loss functions. We omit the discussion for brevity.

5.2. Nonlinear Features

We now consider Rademacher complexity for learning with nonlinear features by DNNs. The following lemma to be proved in Section E.2 gives Rademacher complexity bounds for learning with features by DNNs. We say an activation $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is positive-homogeneous if $\sigma(ax) = a\sigma(x)$ for $a \geq 0$, contractive if $|\sigma(x) - \sigma(x')| \leq |x - x'|$. The ReLU activation function $\sigma(x) = \max\{x, 0\}$ is both positive-homogeneous and contractive.

Proposition 5.5 (Nonlinear representation)**.** *Consider the feature map defined in Lemma 5.1 with $\|\cdot\| = \|\cdot\|_F$ and*

$$\mathcal{V} = \Big\{\mathbf{x} \mapsto \mathbf{v}(x) = \sigma\big(V_L\sigma\big(V_{L-1}\cdots\sigma(V_1\mathbf{x})\big)\big) : \|V_l\|_F \leq B_l, \forall l \in [L]\Big\},$$

where σ is positive-homogeneous, contractive and $\sigma(0) = 0$, and L is the number of layers. Then

$$\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{t\in[d]}\sum_{j\in[n]}\epsilon_{t,j}f_t(\mathbf{x}_j) \leq \sqrt{d}\Lambda B_L B_{L-1}\cdots B_1 \times \Big(16L\big(\sum_{1\leq i<j\leq n}(\mathbf{x}_i^\top\mathbf{x}_j)^2\big)^{\frac{1}{2}} + \sum_{j\in[n]}\|\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}.$$

Remark 5.6. If $d = 1$, the following bound was established in Golowich et al. (2018)

$$\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j f_t(\mathbf{x}_j) = O\Big(\sqrt{L}\Big(\sum_{j\in[n]}\|\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}\prod_{l\in[L]}B_l\Big).$$

Proposition 5.5 extends this bound to the general case $d \in \mathbb{N}$. In particular, if $d = 1$, our result matches the result in Golowich et al. (2018) up to a constant factor. We need to introduce different techniques to handle the difficulty in considering the coupling among different features $\boldsymbol{u}_t^\top\mathbf{v}(x), t \in [d]$, which is reflected by the regularizer on U as $\|U\|_F \leq \Lambda$. Ignoring this coupling would imply a bound with a crude dependency on d. To preserve the coupling, we consider the moment generation function (MGF) of $\sup_{f\in\mathcal{F}}\big(\sum_{t\in[d]}\sum_{j\in[n]}\epsilon_{t,j}f_t(\mathbf{x}_j)\big)^2$, and then reduce it to the MGF of a Rademacher chaos variable $\sum_{1\leq i<j\leq n}\epsilon_i\epsilon_j\mathbf{x}_i^\top\mathbf{x}_j$ by repeated applications of contraction inequalities of Rademacher complexities. A direct application of the analysis in Golowich et al. (2018) show

$$\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j f_t(\mathbf{x}_j) = O\big(d\sqrt{L}\big(\sum_{j\in[n]}\|\mathbf{x}_j\|_2^2\big)^{\frac{1}{2}}\prod_{l\in[L]}B_l\big).$$

As a comparison, our analysis implies a bound with a square-root dependency on d. We will give more details on the comparison of technical analysis in Remark E.6.

Note $\big(\sum_{1\leq i<j\leq n}(\mathbf{x}_i^\top\mathbf{x}_j)^2\big)^{\frac{1}{2}} = O\big(\sum_{j\in[n]}\|\mathbf{x}_j\|_2^2\big)$, from which and Proposition 5.5 we get for nonlinear features that $\mathfrak{C} = O\big(\sqrt{dL}\Lambda(nk)^{\frac{1}{2}}B_x\prod_{l\in[L]}B_l\big)$. The following proposition then follows directly from Theorem 4.9.

Corollary 5.7. *Consider the feature map in Proposition 5.5. Let ℓ be the logistic loss and $\delta \in (0,1)$. With probability at least $1-\delta$ the following inequality holds for all $f \in \mathcal{F}$*

$$L_{un}(f) - \hat{L}_{un}(f) = \widetilde{O}\Big(\frac{GR\sqrt{dL}\Lambda B_x\prod_{l\in[L]}B_l + B\log^{\frac{1}{2}}\frac{1}{\delta}}{\sqrt{n}}\Big).$$

5.3. Generalization for Downstream Classification

In this subsection, we apply the above generalization bounds on unsupervised learning to derive generalization guarantees for a downstream supervised learning task. Similar ideas can be dated back to metric/similarity learning, where one shows that similarity-based learning guarantees a good generalization of the resultant classification (Guo & Ying, 2014; Balcan et al., 2008; Balcan & Blum, 2006). Following Arora et al. (2019), we consider a particular *mean classifier* with rows being the means of the representation of each class, i.e., $\mathbf{x} \mapsto W^\mu f(\mathbf{x})$ with the c-th row of W being the mean μ_c of representations of inputs with label c: $\mu_c := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_c}[f(\mathbf{x})]$. Consider the *average supervised loss*

$$L_{sup}^\mu(f) := \mathbb{E}_{\{c_i\}_{i=1}^{K+1}\sim\rho^{K+1}}\big[L_{sup}(\{c_i\}_{i=1}^{K+1}, W^\mu f)|c_i \neq c_j\big],$$

where we take the expectation over $\mathcal{T} = \{c_i\}_{i=1}^{K+1}$. The following lemma shows that the generalization performance of the mean classifier based on a representation f can be guaranteed in terms of the generalization performance of the representation in unsupervised learning.

Lemma 5.8 (Arora et al. 2019)**.** *There exists a function $\rho : \mathcal{C}^{K+1} \mapsto \mathbb{R}_+$ such that the following inequality holds for any $f \in \mathcal{F}$: $\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\big[\rho(\mathcal{T})L_{sup}^\mu(f)\big] \leq L_{un}(f)$.*

We refer the interested readers to Arora et al. (2019) for the expression of $\rho(\mathcal{T})$, which is independent of n. The following corollaries are immediate applications of Lemma 5.8 and our generalization bounds for unsupervised learning. We omit the proof for brevity.

Corollary 5.9 (Linear representation). *Consider the feature map in Proposition 5.3 with $\|\cdot\| = \|\cdot\|_{2,p}$. Let ℓ_s, ℓ be the logistic loss and $\delta \in (0,1)$. Then with probability at least $1-\delta$ the following inequality holds*

$$\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\Big[\rho(\mathcal{T})L^{\mu}_{sup}(\hat{f})\Big] = \hat{L}_{un}(\hat{f}) + \widetilde{O}\Big(\frac{B\log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}} + \frac{GRB_x \min_{q\geq p}\big\{\Lambda d^{1/q^*}\max(\sqrt{q^*-1},1)\big\}}{\sqrt{n}}\Big).$$

Remark 5.10. If $p \leq (\log d)/(\log d - 1)$, we set $q = (\log d)/(\log d - 1)$, and get $d^{1/q^*}\max(\sqrt{q^*-1},1) = O(\log^{\frac{1}{2}} d)$. In this case, we get a bound with a logarithmic dependency on the number of features. It is possible to extend our discussion to more general norms $\|\cdot\| = \|\cdot\|_{p,q}, p, q \geq 1$ (Kakade et al., 2012).

Corollary 5.11 (Nonlinear representation). *Consider the feature map in Proposition 5.5. Let ℓ_s, ℓ be the logistic loss and $\delta \in (0,1)$. With probability at least $1-\delta$ we have*

$$\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\Big[\rho(\mathcal{T})L^{\mu}_{sup}(\hat{f})\Big] = \hat{L}_{un}(\hat{f}) + \widetilde{O}\Big(\frac{GR\sqrt{dL}\Lambda B_x\prod_{l\in[L]}B_l}{\sqrt{n}} + \frac{B\log^{1/2}(1/\delta)}{\sqrt{n}}\Big).$$

Remark 5.12. If we combine our Rademacher complexity bounds in Section 5 and Eq. (4.3) developed in Arora et al. (2019), we would get generalization bounds for supervised classification with a linear dependency on k. If we combine our complexity bounds and Theorem 4.6, we would get generalization bounds for supervised classification with a square-root dependency on k. These discussions use the Lipschitz continuity of ℓ w.r.t $\|\cdot\|_2$. As a comparison, the use of Lipschitz continuity w.r.t. $\|\cdot\|_\infty$ allows us to derive generalization bounds with a logarithmic dependency on k in Corollary 5.9 and Corollary 5.11. Furthermore, we can improve the bounds $\widetilde{O}(1/\sqrt{n})$ in these corollaries to $\widetilde{O}(1/n)$ in an interpolation setting by applying Theorem 4.13.

Remark 5.13. Note $\rho(\mathcal{T})$ in the above corollaries can grow very fast w.r.t. k, which motivates various studies on the connection between feature learning and downstream classification tasks (Ash et al., 2022; Nozawa & Sato, 2021; Wang et al., 2022; Bao et al., 2022). As a comparison, the main focus of our paper is to improve generalization bounds for pre-train task, which is orthogonal to the analysis in Ash et al. (2022); Nozawa & Sato (2021); Wang et al. (2022); Bao et al. (2022). It should be mentioned that Ash et al. (2022) also studies generalization bounds for pre-train task. Under the assumption $\|f(x)\|_1 \leq R_1$, Ash et al. (2022) shows the following generalization bounds for contrastive learning with ℓ_∞ Lipschitz loss

$$L_{un}(f) - \hat{L}_{un}(f) \lesssim \frac{R_1\sqrt{kd}}{n}\max_{t\in[d]}\max_{\mathbf{x}_1,\ldots,\mathbf{x}_n}\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{j\in[n]}f_t(\mathbf{x}_j),$$

which is worse than our bound by a factor of $\sqrt{k}$. Moreover, this bound uses R_1 while our bound assumes $\|f(x)\|_2 \leq R$. Note R_1 can be larger than R by a factor of $\sqrt{d}$.

It should be mentioned that our improvement on the generalization bounds for pre-train task can be seamlessly combined with the connection on pre-train task and downstream task to get improved generalization bounds for the downstream task. For example, Nozawa & Sato (2021) derives the bound $L^u_{sup}(\hat{f}) \leq \frac{2L_{un}(\hat{f})}{v_{k+1}}$, where v_{k+1} is the probability that the sampled k negative examples contains all classes. We can directly combine this result and our analysis to derive the following bounds on the performance of downstream tasks for learning with linear features

$$L^u_{sup}(\hat{f}) = \frac{2}{v_{k+1}}\Big(\hat{L}_{un}(\hat{f}) + \widetilde{O}\Big(\frac{\log^{1/2}(1/\delta)+\sqrt{d}}{\sqrt{n}}\Big)\Big).$$

6. Conclusion

Motivated by the existing generalization bounds with a crude dependency on the number of negative examples, we present a systematic analysis on the generalization behavior of CRL. We consider three types of loss functions. Our results improve the existing bounds by a factor of $\sqrt{k}$ for ℓ_2 Lipschitz loss, and by a factor of k for ℓ_∞ Lipschitz loss (up to a log factor). We get optimistic bounds for self-bounding Lipschitz loss, which imply fast rates under low noise conditions. We justify the effectiveness of our results with applications to both linear and nonlinear features.

Our analysis based on Rademacher complexities implies algorithm-independent bounds. It would be interesting to develop algorithm-dependent bounds to understand the interaction between optimization and generalization. For ℓ_∞ loss, our bound still enjoys a logarithmic dependency on k. It would be interesting to study whether this logarithmic dependency can be removed in further study.

Acknowledgments

We thank the reviewers and area chair for their constructive comments. Part of the work was done when Yunwen Lei was with the Department of Mathematics, Hong Kong Baptist University. The work of Tianbao Yang was partially supported by NSF Career Award #1844403, NSF Program #2110545, and NSF-Amazon Joint Program #2147253. The work of Yiming Ying was partially supported by NSF (DMS-2110836, IIS-2103450, and IIS-2110546). The work of Ding-Xuan Zhou was partially supported by InnoHK initiative, The Government of the HKSAR, Laboratory for AI-Powered Financial Technologies, Research Grant Council of Hong Kong [Projects # CityU 11308121, N_CityU102/20, C1013-21GF], and NSFC [Project No. 12061160462].

References

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*, volume 9. 1999.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pp. 9904–9923, 2019.

Ash, J., Goel, S., Krishnamurthy, A., and Misra, D. Investigating the role of negatives in contrastive representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 7187–7209. PMLR, 2022.

Balcan, M.-F. and Blum, A. On a theory of learning with similarity functions. In *international Conference on Machine Learning*, pp. 73–80, 2006.

Balcan, M.-F., Blum, A., and Srebro, N. Improved guarantees for learning via similarity functions. In *Conference on Learning Theory*, pp. 287–298, 2008.

Bao, H., Nagano, Y., and Nozawa, K. On the surrogate gap between contrastive and supervised losses. In *International Conference on Machine Learning*, pp. 1585–1606. PMLR, 2022.

Bartlett, P. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cao, Q., Guo, Z.-C., and Ying, Y. Generalization bounds for metric and similarity learning. *Machine Learning*, 102 (1):115–132, 2016.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Cucker, F. and Zhou, D.-X. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, 2007.

De la Pena, V. and Giné, E. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Foster, D. J. and Rakhlin, A. ℓ_∞ vector contraction for rademacher complexity. *arXiv preprint arXiv:1911.06468*, 2019.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Guermeur, Y. Lp-norm sauer–shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017.

Guo, Z.-C. and Ying, Y. Guaranteed classification via regularized similarity learning. *Neural Computation*, 26(3): 497–522, 2014.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Hénaff, O. J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S. A., and Van Den Oord, A. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192, 2020.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Lei, Y., Dogan, U., Binder, A., and Kloft, M. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems*, pp. 2026–2034, 2015.

Lei, Y., Dogan, Ü., Zhou, D.-X., and Kloft, M. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5): 2995–3021, 2019.

Lust-Piquard, F. and Pisier, G. Non commutative khintchine and paley inequalities. *Arkiv för Matematik*, 29(1):241–260, 1991.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17, 2016.

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT press, 2012.

Mustafa, W., Lei, Y., and Kloft, M. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pp. 16174–16196. PMLR, 2022.

Nozawa, K. and Sato, I. Understanding negative samples in instance discriminative self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:5784–5797, 2021.

Nozawa, K., Germain, P., and Guedj, B. PAC-bayesian contrastive unsupervised representation learning. In *Uncertainty in Artificial Intelligence*, pp. 21–30. PMLR, 2020.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Reeve, H. and Kaban, A. Optimistic bounds for multi-output learning. In *International Conference on Machine Learning*, pp. 8030–8040. PMLR, 2020.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, volume 162, pp. 19250–19286, 2022.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pp. 2199–2207, 2010.

Tewari, A. and Chaudhuri, S. Generalization error bounds for learning to rank: Does the length of document lists matter? In *International Conference on Machine Learning*, pp. 315–323. PMLR, 2015.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European Conference on Computer Vision*, pp. 776–794. Springer, 2020a.

Tian, Y., Yu, L., Chen, X., and Ganguli, S. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020b.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021a.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research*, 22:281–1, 2021b.

Tropp, J. A. The expected norm of a sum of independent random matrices: An elementary approach. In *High Dimensional Probability VII: The Cargèse Volume*, pp. 173–202. Springer, 2016.

Tsai, Y.-H. H., Wu, Y., Salakhutdinov, R., and Morency, L.-P. Demystifying self-supervised learning: An information-theoretical framework. *arXiv preprint arXiv:2006.05576*, 2, 2020.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Wang, Y., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022.

Ying, Y. and Campbell, C. Rademacher chaos complexities for learning the kernel problem. *Neural computation*, 22 (11):2858–2886, 2010.

Yuan, Z., Wu, Y., Qiu, Z., Du, X., Zhang, L., Zhou, D., and Yang, T. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In *Proceedings of International Conference of Machine Learning*, 2022.

Zhou, D.-X. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Appendix for "Generalization Analysis for Contrastive Representation Learning"

A. Effect on the Preservation of Coupling in Lemma 4.3

In this section, we show that Lemma 4.3 can improve Eq. (4.1) by a factor of $\sqrt{d}$ due to the ability in preserving the coupling among different coordinates in the features. To this aim, we introduce the following lemma on the spectral norm of random matrices.

Lemma A.1 (Tropp 2016). *Consider an independent family $S_1, \ldots, S_n$ of random $d \times d$ matrices with $\mathbb{E}[S_i] = 0$ for each i, and define $Z = \sum_{i\in[n]} S_i$. Denote $C(d) = 4(1 + 2\lceil \log d \rceil)$. Then*

$$\mathbb{E}[\|Z\|_{op}] \leq \sqrt{C(d)} \max\Big\{ \big\|\mathbb{E}[ZZ^\top]\big\|_{op}^{\frac{1}{2}}, \big\|\mathbb{E}[Z^\top Z]\big\|_{op}^{\frac{1}{2}} \Big\} + C(d)\big(\mathbb{E}\max_i \|S_i\|_{op}^2\big)^{\frac{1}{2}}.$$

For simplicity, let us consider linear features, i.e.,

$$\mathcal{F} = \Big\{ f(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_d(\mathbf{x}))^\top : \sum_{t\in[d]} \|\mathbf{w}_t\|_2^2 \leq R_w^2 \Big\},$$

where $f_t(\mathbf{x}) = \langle \mathbf{w}_t, \mathbf{x} \rangle$ and $\|\mathbf{x}\|_2 \leq 1$. In this case, we have

$$\|f(\mathbf{x})\|_\infty = \max_{t\in[d]} |f_t(\mathbf{x})| \leq \max_{t\in[d]} \|\mathbf{w}_t\|_2 \|\mathbf{x}\|_2 \leq R_w$$

and

$$\|f(\mathbf{x})\|_2 = \Big(\sum_{t\in[d]} f_t^2(\mathbf{x}) \Big)^{\frac{1}{2}} \leq \Big(\sum_{t\in[d]} \|\mathbf{w}_t\|_2^2 \|\mathbf{x}\|_2^2 \Big)^{\frac{1}{2}} \leq R_w.$$

Eq. (4.1) implies

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{j\in[n]} \epsilon_j f(\mathbf{x}_j)^\top f(\mathbf{x}_j^+) &\leq d \max_{t\in[d]} \mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{n}} \Big[\sup_{f\in\mathcal{F}} \sum_{j\in[n]} \epsilon_j f_t(\mathbf{x}_j) f_t(\mathbf{x}_j^+) \Big] \\
&= d \max_{t\in[d]} \mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{n}} \Big[\sup_{f\in\mathcal{F}} \sum_{j\in[n]} \epsilon_j \mathbf{w}_t^\top \mathbf{x}_j^+ \mathbf{x}_j^\top \mathbf{w}_t \Big] \\
&= d \max_{t\in[d]} \mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{n}} \Big[\sup_{\|\mathbf{w}_t\|_2\leq R_w} \mathbf{w}_t^\top \Big(\sum_{j\in[n]} \epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top \Big) \mathbf{w}_t \Big] \\
&= d R_w^2 \mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^\mathbf{n}} \Big\| \sum_{j\in[n]} \epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top \Big\|_{\text{op}},
\end{aligned}$$

where $\|\cdot\|_{\text{op}}$ denotes the spectral operator of a matrix. Let $Z = \sum_{j\in[n]} \epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top$. It is clear that

$$\mathbb{E}\Big[\Big(\sum_{j\in[n]} \epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top \Big) \Big(\sum_{j\in[n]} \epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top \Big)^\top \Big] = \sum_{j\in[n]} \mathbf{x}_j^+ \mathbf{x}_j^\top \mathbf{x}_j (\mathbf{x}_j^+)^\top \preceq \sum_{j\in[n]} \mathbf{x}_j^+ (\mathbf{x}_j^+)^\top$$

and

$$\mathbb{E}\Big[\Big(\sum_{j\in[n]} \epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top \Big)^\top \Big(\sum_{j\in[n]} \epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top \Big) \Big] = \sum_{j\in[n]} \mathbf{x}_j (\mathbf{x}_j^+)^\top \mathbf{x}_j^+ \mathbf{x}_j^\top \preceq \sum_{j\in[n]} \mathbf{x}_j (\mathbf{x}_j)^\top.$$

Therefore, we have

$$\max\Big\{ \big\|\mathbb{E}[ZZ^\top]\big\|_{\text{op}}^{\frac{1}{2}}, \big\|\mathbb{E}[Z^\top Z]\big\|_{\text{op}}^{\frac{1}{2}} \Big\} \leq \sqrt{n}.$$

Furthermore, we have

$$\mathbb{E}\max_j \|\epsilon_j \mathbf{x}_j^+ \mathbf{x}_j^\top\|_{\text{op}}^2 \leq 1.$$

It then follows from Lemma A.1 that

$$\mathbb{E}_{\epsilon\sim\{\pm\mathbf{1}\}^{\mathbf{n}}}\Big\|\sum_{j\in[n]}\epsilon_j\mathbf{x}_j^+\mathbf{x}_j^\top\Big\|_{\text{op}}\leq\sqrt{nC(d)}+C(d).$$

We can combine the above discussions together to derive the following inequality based on Eq. (4.1)

$$\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j f(\mathbf{x}_j)^\top f(\mathbf{x}_j^+)\leq dR_w^2\Big(\sqrt{nC(d)}+C(d)\Big). \tag{A.1}$$

As a comparison, the inequality below Eq. (4.1) implies the following inequality via our approach

$$\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j f(\mathbf{x}_j)^\top f(\mathbf{x}_j^+)\lesssim R_w\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\epsilon_{j,t}f_t(\mathbf{x}_j).$$

Furthermore, there holds

$$\begin{aligned}\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\epsilon_{j,t}f_t(\mathbf{x}_j)&=\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\epsilon_{j,t}\langle\mathbf{w}_t,\mathbf{x}_j\rangle=\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{t\in[d]}\Big\langle\mathbf{w}_t,\sum_{j\in[n]}\epsilon_{j,t}\mathbf{x}_j\Big\rangle\\&\leq\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{t\in[d]}\|\mathbf{w}_t\|_2\Big\|\sum_{j\in[n]}\epsilon_{j,t}\mathbf{x}_j\Big\|_2\leq R_w\mathbb{E}_\epsilon\Big(\sum_{t\in[d]}\|\sum_{j\in[n]}\epsilon_{j,t}\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}\\&\leq R_w\Big(\sum_{t\in[d]}\mathbb{E}_\epsilon\|\sum_{j\in[n]}\epsilon_{j,t}\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}=R_w\Big(\sum_{t\in[d]}\mathbb{E}_\epsilon\sum_{j\in[n]}\|\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}\leq R_w\sqrt{nd},\end{aligned}$$

where we have used Cauchy-Schwartz's inequality in the first inequality and the Jensen's inequality in the second inequality. Therefore, our analysis implies

$$\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j f(\mathbf{x}_j)^\top f(\mathbf{x}_j^+)\leq R_w^2\sqrt{nd}.$$

It is clear our analysis improves (A.1) based on existing analysis by a factor of $\sqrt{d}$ in this specific problem.

B. Proof of Theorem 4.6

To prove Theorem 4.6, we first prove Lemma 4.3 by the following vector-contraction lemma on Rademacher complexities.

Lemma B.1 (Maurer 2016). *Let $S=\{\mathbf{z}_j\}_{j=1}^n\in\mathcal{Z}^n$. Let $\mathcal{F}'$ be a class of functions $f':\mathcal{Z}\mapsto\mathbb{R}^d$ and $h:\mathbb{R}^d\mapsto\mathbb{R}$ be G-Lipschitz w.r.t. ℓ_2-norm. Then*

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\Big[\sup_{f'\in\mathcal{F}'}\sum_{j\in[n]}\epsilon_j(h\circ f')(\mathbf{z}_j)\Big]\leq\sqrt{2}G\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}}\Big[\sup_{f'\in\mathcal{F}'}\sum_{j\in[n]}\sum_{t\in[d]}\epsilon_{j,t}f_t'(\mathbf{z}_j)\Big].$$

In Section F, we will provide an extension of the above lemma.

Proof of Lemma 4.3. Let $f':\mathcal{X}^3\mapsto\mathbb{R}^{3d}$ be defined as

$$f'(\mathbf{x},\mathbf{x}^+,\mathbf{x}^-)=\big(f(\mathbf{x}),f(\mathbf{x}^+),f(\mathbf{x}^-)\big)\in\mathbb{R}^{3d}$$

and $h:\mathbb{R}^{3d}\mapsto\mathbb{R}$ be defined as

$$h(\mathbf{y},\mathbf{y}^+,\mathbf{y}^-)=\mathbf{y}^\top(\mathbf{y}^+-\mathbf{y}^-),\quad\mathbf{y},\mathbf{y}^+,\mathbf{y}^-\in\mathbb{R}^d.$$

Then it is clear that

$$f(\mathbf{x})^\top(f(\mathbf{x}^+)-f(\mathbf{x}^-))=h\circ f'(\mathbf{x},\mathbf{x}^+,\mathbf{x}^-).$$

Furthermore, for any $\mathbf{y}_1,\mathbf{y}_1^+,\mathbf{y}_1^-,\mathbf{y}_2,\mathbf{y}_2^+,\mathbf{y}_2^-$ with Euclidean norm less than or equal to R, we have

$$\begin{aligned}&h(\mathbf{y}_1,\mathbf{y}_1^+,\mathbf{y}_1^-)-h(\mathbf{y}_2,\mathbf{y}_2^+,\mathbf{y}_2^-)=\mathbf{y}_1^\top(\mathbf{y}_1^+-\mathbf{y}_1^-)-\mathbf{y}_2^\top(\mathbf{y}_2^+-\mathbf{y}_2^-)\\&=\mathbf{y}_1^\top(\mathbf{y}_1^+-\mathbf{y}_1^-)-\mathbf{y}_1^\top(\mathbf{y}_2^+-\mathbf{y}_2^-)+\mathbf{y}_1^\top(\mathbf{y}_2^+-\mathbf{y}_2^-)-\mathbf{y}_2^\top(\mathbf{y}_2^+-\mathbf{y}_2^-)\\&=\mathbf{y}_1^\top(\mathbf{y}_1^+-\mathbf{y}_1^--\mathbf{y}_2^++\mathbf{y}_2^-)+(\mathbf{y}_1-\mathbf{y}_2)^\top(\mathbf{y}_2^+-\mathbf{y}_2^-).\end{aligned}$$

It then follows from the elementary inequality $(a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2$ that

$$\begin{aligned}\big|h(\mathbf{y}_1, \mathbf{y}_1^+, \mathbf{y}_1^-) - h(\mathbf{y}_2, \mathbf{y}_2^+, \mathbf{y}_2^-)\big|^2 \leq \\ 2(1+p)\|\mathbf{y}_1\|^2\|\mathbf{y}_1^+ - \mathbf{y}_2^+\|^2 + 2(1+p)\|\mathbf{y}_1\|^2\|\mathbf{y}_1^- - \mathbf{y}_2^-\|_2^2 + (1+1/p)\|\mathbf{y}_2^+ - \mathbf{y}_2^-\|_2^2\|\mathbf{y}_1 - \mathbf{y}_2\|_2^2.\end{aligned}$$

We can choose $p = 2$ and get

$$\begin{aligned}\big|h(\mathbf{y}_1, \mathbf{y}_1^+, \mathbf{y}_1^-) - h(\mathbf{y}_2, \mathbf{y}_2^+, \mathbf{y}_2^-)\big|^2 &\leq 6R^2\big(\|\mathbf{y}_1^+ - \mathbf{y}_2^+\|^2 + \|\mathbf{y}_1^- - \mathbf{y}_2^-\|_2^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2\big) \\ &= 6R^2\|(\mathbf{y}_1, \mathbf{y}_1^+, \mathbf{y}_1^-) - (\mathbf{y}_2, \mathbf{y}_2^+, \mathbf{y}_2^-)\|_2^2.\end{aligned}$$

This shows that h is $\sqrt{6}R$-Lipschitz continuous w.r.t. $\|\cdot\|_2$. We can apply Lemma B.1 to derive

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\Big[\sup_{f'\in\mathcal{F}'}\sum_{j\in[n]}\epsilon_j(h\circ f')(\mathbf{z}_j)\Big] \\ &\qquad\leq \sqrt{12}R\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n\times\{\pm1\}^d\times\{\pm1\}^3}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{t\in[d]}\Big(\epsilon_{j,t,1}f_t(\mathbf{x}_j) + \epsilon_{j,t,2}f_t(\mathbf{x}_j^+) + \epsilon_{j,t,3}f_t(\mathbf{x}_j^-)\Big)\Big].\end{aligned}$$

The proof is completed. □

The following standard lemma gives generalization error bounds in terms of Rademacher complexities.

Lemma B.2 (Mohri et al. 2012)**.** *Let $\widetilde{\mathcal{G}}$ be a function class and $\widetilde{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. If for any $g \in \widetilde{\mathcal{G}}$ we have $g(\mathbf{z}) \in [0, B]$, then for any $\delta \in (0,1)$ the following inequality holds with probability (w.r.t. $\widetilde{S}$) at least $1-\delta$*

$$\mathbb{E}[g(\mathbf{z})] \leq \frac{1}{n}\sum_{i=1}^n g(\mathbf{z}_i) + 2\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{G}}) + 3B\sqrt{\frac{\log(2/\delta)}{2n}}, \quad \forall g \in \widetilde{\mathcal{G}}.$$

Proof of Theorem 4.6. According to the G_2-Lipschitz continuity of ℓ w.r.t. ℓ_2-norm and Lemma B.1, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\epsilon_j\ell\big(\big\{f(\mathbf{x}_j)^\top(f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-))\big\}_{i\in[k]}\big)\Big] \\ &\qquad\leq \sqrt{2}G_2\mathbb{E}_{\{\boldsymbol{\epsilon}\}\sim\{\pm1\}^{nk}}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{i\in[k]}\epsilon_{j,i}f(\mathbf{x}_j)^\top(f(\mathbf{x}_j^\top) - f(\mathbf{x}_{ji}^-))\Big].\end{aligned}$$

According to Lemma 4.3, we further get

$$\begin{aligned}&\mathbb{E}_{\{\boldsymbol{\epsilon}\}\sim\{\pm1\}^{nk}}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{i\in[k]}\epsilon_{j,i}f(\mathbf{x}_j)^\top(f(\mathbf{x}_j^\top) - f(\mathbf{x}_{ji}^-))\Big] \leq \\ &\qquad\sqrt{12}R\mathbb{E}_{\{\boldsymbol{\epsilon}\}\sim\{\pm1\}^{3nkd}}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\sum_{j\in[n]}\sum_{i\in[k]}\sum_{t\in[d]}\Big(\epsilon_{j,i,t,1}f_t(\mathbf{x}_j) + \epsilon_{j,i,t,2}f_t(\mathbf{x}_j^+) + \epsilon_{j,i,t,3}f_t(\mathbf{x}_{ji}^-)\Big)\Big].\end{aligned}$$

We can combine the above two inequalities to get the Rademacher complexity bounds.

We now turn to the generalization bounds. Applying Lemma B.2, with probability at least $1-\delta$ the following inequality holds with probability at least $1-\delta$

$$L_{un}(f) \leq \hat{L}_{un}(f) + 2\mathfrak{R}_S(\mathcal{G}) + 3B\sqrt{\frac{\log(2/\delta)}{2n}}, \quad \forall f \in \mathcal{F}.$$

The stated bound on generalization errors then follows by plugging the Rademacher complexity bounds into the above bound. The proof is completed. □

C. Proof of Theorem 4.9

We first introduce several complexity measures such as covering numbers and fat-shattering dimension (Alon et al., 1997; Anthony & Bartlett, 1999; Cucker & Zhou, 2007; Zhou, 2002).

Definition C.1 (Covering number)**.** Let $\widetilde{S} := \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \in \mathcal{Z}^n$. Let $\widetilde{\mathcal{F}}$ be a class of real-valued functions defined over a space $\mathcal{Z}$. For any $\epsilon > 0$ and $p \geq 1$, the empirical ℓ_p-norm covering number $\mathcal{N}_p(\epsilon, \widetilde{\mathcal{F}}, \widetilde{S})$ with respect to $\widetilde{S}$ is defined as the smallest number m of a collection of vectors $\mathbf{v}^1, \ldots, \mathbf{v}^m \in \{(f(\mathbf{z}_1), \ldots, f(\mathbf{z}_n)) : f \in \widetilde{\mathcal{F}}\}$ such that

$$\sup_{f \in \widetilde{\mathcal{F}}} \min_{j \in [m]} \Big(\frac{1}{n} \sum_{i \in [n]} |f(\mathbf{z}_i) - \mathbf{v}_i^j|^p \Big)^{\frac{1}{p}} \leq \epsilon,$$

where $\mathbf{v}_i^j$ is the i-th component of the vector $\mathbf{v}^j$. In this case, we call $\{\mathbf{v}^1, \ldots, \mathbf{v}^m\}$ an (ϵ, ℓ_p)-cover of $\widetilde{\mathcal{F}}$ with respect to $\widetilde{S}$.

Definition C.2 (Fat-Shattering Dimension)**.** Let $\widetilde{\mathcal{F}}$ be a class of real-valued functions defined over a space $\widetilde{\mathcal{Z}}$. We define the fat-shattering dimension $\mathrm{fat}_\epsilon(\widetilde{\mathcal{F}})$ at scale $\epsilon > 0$ as the largest $D \in \mathbb{N}$ such that there exist D points $\mathbf{z}_1, \ldots, \mathbf{z}_D \in \widetilde{\mathcal{Z}}$ and witnesses $s_1, \ldots, s_D \in \mathbb{R}$ satisfying: for any $\delta_1, \ldots, \delta_D \sim \{\pm 1\}$ there exists $f \in \widetilde{\mathcal{F}}$ with

$$\delta_i (f(\mathbf{z}_i) - s_i) \geq \epsilon/2, \qquad \forall i \in [D].$$

To prove Theorem 4.8, we need to introduce the following lemma on Rademacher complexity, fat-shattering dimension and covering numbers. Part (a) shows that the covering number can be bounded by fat-shattering dimension (see, e.g., Theorem 12.8 in Anthony & Bartlett (1999)). Part (b) shows that the fat-shattering dimension can be controlled by the worst-case Rademacher complexity, which was developed in Srebro et al. (2010). Part (c) is a discretization of the chain integral to control Rademacher complexity by covering numbers (Srebro et al., 2010), which can be found in Guermeur (2017). Let e be the base of the natural logarithms.

Lemma C.3. *Let $\widetilde{S} := \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \subseteq \widetilde{\mathcal{Z}}$. Let $\widetilde{\mathcal{F}}$ be a class of real-valued functions defined over a space $\widetilde{\mathcal{Z}}$.*

(a) If functions in $\widetilde{\mathcal{F}}$ take values in $[-B, B]$, then for any $\epsilon > 0$ with $\mathrm{fat}_\epsilon(\widetilde{\mathcal{F}}) < n$ we have

$$\log \mathcal{N}_\infty(\epsilon, \widetilde{\mathcal{F}}, \widetilde{S}) \leq 1 + \mathrm{fat}_{\epsilon/4}(\widetilde{\mathcal{F}}) \Big(\log_2 \frac{4eBn}{\epsilon \mathrm{fat}_{\epsilon/4}(\widetilde{\mathcal{F}})} \Big) \Big(\log \frac{16B^2 n}{\epsilon^2} \Big).$$

(b) For any $\epsilon > \mathfrak{R}_{\widetilde{\mathcal{Z}},n}(\widetilde{\mathcal{F}})$, we have $\mathrm{fat}_\epsilon(\widetilde{\mathcal{F}}) < \frac{4n}{\epsilon^2} \mathfrak{R}^2_{\widetilde{\mathcal{Z}},n}(\widetilde{\mathcal{F}})$.

(c) Let $(\epsilon_j)_{j=0}^\infty$ be a monotone sequence decreasing to 0 and any $(a_1, \ldots, a_n) \in \mathbb{R}^n$. If $\epsilon_0 \geq \sqrt{n^{-1} \sup_{f \in \widetilde{\mathcal{F}}} \sum_{i=1}^n \big(f(\mathbf{z}_i) - a_i\big)^2}$, then for any non-negative integer N we have

$$\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}}) \leq 2 \sum_{j=1}^N (\epsilon_j + \epsilon_{j-1}) \sqrt{\frac{\log \mathcal{N}_\infty(\epsilon_j, \widetilde{\mathcal{F}}, \widetilde{S})}{n}} + \epsilon_N. \tag{C.1}$$

According to Part (a) of Lemma C.3, the following inequality holds for any $\epsilon \in (0, 2B]$ (the case $\mathrm{fat}_{\epsilon/4}(\widetilde{\mathcal{F}}) = 0$ is trivial since in this case we have $\mathcal{N}_\infty(\epsilon, \widetilde{\mathcal{F}}, \widetilde{S}) = 1$, and otherwise we have $\mathrm{fat}_{\epsilon/4}(\widetilde{\mathcal{F}}) \geq 1$)

$$\log \mathcal{N}_\infty(\epsilon, \widetilde{\mathcal{F}}, \widetilde{S}) \leq 1 + \mathrm{fat}_{\epsilon/4}(\widetilde{\mathcal{F}}) \log_2^2 \frac{8eB^2 |\widetilde{S}|}{\epsilon^2}. \tag{C.2}$$

We follow the arguments in Lei et al. (2019) to prove Theorem 4.8.

Proof of Theorem 4.8. We first relate the empirical ℓ_∞-covering number of $\mathcal{F}$ w.r.t. $S = \{(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_{j1}^-, \mathbf{x}_{j2}^-, \ldots, \mathbf{x}_{jk}^-) : j \in [n]\}$ to the empirical ℓ_∞-covering number of $\mathcal{H}$ w.r.t. $S_\mathcal{H}$. Let

$$\Big\{ \mathbf{r}^m = \big(r_{1,1}^m, r_{1,2}^m, \ldots, r_{1,k}^m, \ldots, r_{n,1}^m, r_{n,2}^m, \ldots, r_{n,k}^m\big) : m \in [N] \Big\}$$

be an $(\epsilon/G, \ell_\infty)$-cover of $\mathcal{H}$ w.r.t. $S_\mathcal{H}$. Recall that

$$h_f(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = f(\mathbf{x})^\top \big(f(\mathbf{x}^+) - f(\mathbf{x}^-)\big). \tag{C.3}$$

Then, by the definition of ℓ_∞-cover we know for any $f \in \mathcal{F}$ we can find $m \in [N]$ such that

$$\max_{j \in [n]} \max_{i \in [k]} \big| h_f(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_{ji}^-) - r_{j,i}^m \big| \leq \epsilon/G.$$

By the Lipschitz continuity of ℓ, we then get

$$\begin{aligned}
&\max_{j\in[n]}\left|\ell\big(\big\{f(\mathbf{x}_j)^\top\big(f(\mathbf{x}_j^+)-f(\mathbf{x}_{ji}^-)\big)\big\}_{i=1}^k\big)-\ell\big(\{r_{j,i}^m\}_{i=1}^k\big)\right| \\
&\le G\big\|\big(f(\mathbf{x}_j)^\top\big(f(\mathbf{x}_j^+)-f(\mathbf{x}_{ji}^-)\big)\big)_{i=1}^k-\big(r_{j,i}^m\big)_{i=1}^k\big\|_\infty = G\big\|\big(h_f(\mathbf{x}_j,\mathbf{x}_j^+,\mathbf{x}_{ji}^-)\big)_{i=1}^k-\big(r_{j,i}^m\big)_{i=1}^k\big\|_\infty \\
&\le G\epsilon/G=\epsilon.
\end{aligned}$$

This shows that $\big\{\big(\ell\big(\{r_{1,i}^m\}_{i=1}^k\big),\ell\big(\{r_{2,i}^m\}_{i=1}^k\big),\ldots,\ell\big(\{r_{n,i}^m\}_{i=1}^k\big)\big):m\in[N]\big\}$ is an (ϵ,ℓ_∞)-cover of $\mathcal{G}$ w.r.t. S and therefore

$$\mathcal{N}_\infty(\epsilon,\mathcal{G},S)\le\mathcal{N}_\infty(\epsilon/G,\mathcal{H},S_{\mathcal{H}}). \tag{C.4}$$

Since we consider empirical covering number of $\mathcal{F}$ w.r.t. S, we can assume functions in $\mathcal{H}$ are defined over $S_{\mathcal{H}}$. For simplicity, we denote $\mathfrak{R}_{nk}(\mathcal{H}):=\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H})$. We now control $\mathcal{N}_\infty(\epsilon/G,\mathcal{H},S_{\mathcal{H}})$ by Rademacher complexities of $\mathcal{H}$. For any $\epsilon>2\mathfrak{R}_{nk}(\mathcal{H})$, it follows from Part (b) of Lemma C.3 that

$$\mathrm{fat}_\epsilon(\mathcal{H})\le\frac{4nk}{\epsilon^2}\mathfrak{R}^2_{S_{\mathcal{H}},nk}(\mathcal{H})\le nk. \tag{C.5}$$

Note for any $f\in\mathcal{F}$, we have $f(\mathbf{x})^\top(f(\mathbf{x}^+)-f(\mathbf{x}^-))\in[-2R^2,2R^2]$. It then follows from Eq. (C.2) and Eq. (C.5) that (replace B by $2R^2$)

$$\begin{aligned}
\log\mathcal{N}_\infty(\epsilon,\mathcal{H},S_{\mathcal{H}})&\le 1+\mathrm{fat}_{\epsilon/4}(\mathcal{H})\log^2(32eR^4nk/\epsilon^2) \\
&\le 1+\frac{64nk\mathfrak{R}^2_{nk}(\mathcal{H})}{\epsilon^2}\log^2(32eR^4nk/\epsilon^2),\quad \epsilon\in(0,4R^2].
\end{aligned}$$

We can combine the above inequality and Eq. (C.4) to derive the following inequality for any $2G\mathfrak{R}_{nk}(\mathcal{H})\le\epsilon\le 4GR^2$

$$\log\mathcal{N}_\infty(\epsilon,\mathcal{G},S)\le 1+\frac{64nkG^2\mathfrak{R}^2_{nk}(\mathcal{H})}{\epsilon^2}\log^2(32eR^4G^2nk/\epsilon^2). \tag{C.6}$$

Let $\epsilon_N=24G\max\big\{\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H}),n^{-\frac{1}{2}}\big\}$,

$$\epsilon_j=2^{N-j}\epsilon_N,\quad j=0,\ldots,N-1,$$

where

$$N=\left\lceil\log_2\frac{2GR^2}{24G\max\big\{\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H}),n^{-\frac{1}{2}}\big\}}\right\rceil.$$

It is clear from the definition that

$$\epsilon_0\ge 2GR^2\ge\epsilon_0/2.$$

The Lipschitz continuity of ℓ implies

$$\ell((\{h_f(\mathbf{x},\mathbf{x}^+,\mathbf{x}_i^-)\}_{i\in[k]}))-\ell((0,0,\ldots,0))\le G\|h_f(\mathbf{x},\mathbf{x}^+,\mathbf{x}^-)\|_\infty\le 2R^2G.$$

According to the above inequality and Part (c) of Lemma C.3, we know (note $\epsilon_N\ge 2G\mathfrak{R}_{nk}(\mathcal{H})$ and therefore Eq. (C.6) holds for $\epsilon=\epsilon_j,j=1,\ldots,N$)

$$\begin{aligned}
\mathfrak{R}_S(\mathcal{G})&\le 2\sum_{j=1}^N(\epsilon_j+\epsilon_{j-1})\sqrt{\frac{\log\mathcal{N}_\infty(\epsilon_j,\mathcal{G},S)}{n}}+\epsilon_N \\
&\le 2n^{-\frac{1}{2}}\sum_{j=1}^N(\epsilon_j+\epsilon_{j-1})+\frac{16G\sqrt{nk}\mathfrak{R}_{nk}(\mathcal{H})}{\sqrt{n}}\sum_{j=1}^N\frac{(\epsilon_j+\epsilon_{j-1})\log(32eR^4G^2nk/\epsilon_j^2)}{\epsilon_j}+\epsilon_N \\
&\le 6\epsilon_0n^{-\frac{1}{2}}+\epsilon_N+\frac{48G\sqrt{nk}\mathfrak{R}_{nk}(\mathcal{H})}{\sqrt{n}}\sum_{j=1}^N\log(32eR^4G^2nk/\epsilon_j^2) \\
&\le 24GR^2n^{-\frac{1}{2}}+\epsilon_N+48GN\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H})\log(32eR^4G^2nk)+48G\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H})\sum_{j=1}^N\log(1/\epsilon_j^2).
\end{aligned}$$

According to the definition of ϵ_k, we know

$$\begin{aligned}\sum_{j=1}^{N}\log(1/\epsilon_j^2) &= \sum_{j=1}^{N}\log(2^{2j}/\epsilon_0^2) = \sum_{j=1}^{N}\log(1/\epsilon_0^2) + \log 4\cdot\sum_{j=1}^{N} j = N\log(1/\epsilon_0^2) + \frac{N(N+1)\log 4}{2} \\ &= N\Big(\log 1/\epsilon_0^2 + (N+1)\log 2\Big) = N\log 2^{N+1}/\epsilon_0^2 = N\log\Big(\frac{1}{\epsilon_N}\frac{2}{\epsilon_0}\Big) \leq N\log\Big(\frac{1}{\epsilon_N}\frac{2}{2GR^2}\Big) \\ &\leq N\log\Big(\frac{\sqrt{n}}{24G}\frac{1}{GR^2}\Big) = N\log\frac{\sqrt{n}}{24G^2R^2}.\end{aligned}$$

We can combine the above two inequalities together to get

$$\begin{aligned}\mathfrak{R}_S(\mathcal{G}) &\leq 24GR^2n^{-\frac{1}{2}} + \epsilon_N + 48NG\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H})\Big(\log(32eR^4G^2nk) + \log\frac{\sqrt{n}}{24G^2R^2}\Big) \\ &\leq 24GR^2n^{-\frac{1}{2}} + \epsilon_N + 48G\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H})\Big(\log(32eR^4G^2nk) + \log\frac{\sqrt{n}}{24G^2R^2}\Big)\Big\lceil\log_2\frac{2GR^2\sqrt{n}}{24G}\Big\rceil \\ &\leq 24GR^2n^{-\frac{1}{2}} + \epsilon_N + 48G\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H})\Big(\log(4R^2n^{\frac{3}{2}}k)\Big)\Big\lceil\log_2\frac{R^2\sqrt{n}}{12}\Big\rceil,\end{aligned}$$

where we have used the definition of N and $32e/24 \leq 4$. The proof is completed. □

Proof of Theorem 4.9. Applying Lemma B.2, with probability at least $1-\delta$ the following inequality holds with probability at least $1-\delta$

$$L_{un}(f) \leq \hat{L}_{un}(f) + 2\mathfrak{R}_S(\mathcal{G}) + 3B\sqrt{\frac{\log(2/\delta)}{2n}}, \quad \forall f \in \mathcal{F}.$$

According to Theorem 4.8 and Lemma 4.3, we know

$$\begin{aligned}\mathfrak{R}_S(\mathcal{G}) &\leq 48G(R^2+1)n^{-\frac{1}{2}} + 48G\sqrt{k}\mathfrak{R}_{nk}(\mathcal{H})\Big(1+\log(R^2n^{\frac{3}{2}}k)\Big\lceil\log_2\frac{R^2\sqrt{n}}{12}\Big\rceil\Big) \\ &\leq 48G(R^2+1)n^{-\frac{1}{2}} + \frac{48\sqrt{12}GR\sqrt{k}}{nk}\Big(1+\log(R^2n^{\frac{3}{2}}k)\Big\lceil\log_2\frac{R^2\sqrt{n}}{12}\Big\rceil\Big)\mathfrak{C}.\end{aligned}$$

We can combine the above two inequalities together and derive the stated bound. The proof is completed. □

D. Proof of Theorem 4.13

To prove Theorem 4.13, we introduce the following lemma on generalization error bounds in terms of local Rademacher complexities (Reeve & Kaban, 2020).

Lemma D.1 (Reeve & Kaban 2020)**.** *Consider a function class $\mathcal{G}$ of functions mapping $\mathcal{Z}$ to $[0,b]$. For any $\widetilde{S} = \{\mathbf{z}_i : i \in [n]\}$ and $g \in \mathcal{G}$, let $\hat{\mathbb{E}}_{\widetilde{S}}[g] = \frac{1}{n}\sum_{i\in[n]} g(\mathbf{z}_i)$. Assume for any $\widetilde{S} \in \mathcal{Z}^n$ and $r > 0$, we have*

$$\mathfrak{R}_{\widetilde{S}}\big(\{g \in \mathcal{G} : \hat{\mathbb{E}}_{\widetilde{S}}[g] \leq r\}\big) \leq \phi_n(r),$$

where $\phi_n : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is non-decreasing and $\phi_n(r)/\sqrt{r}$ is non-increasing. Let $\hat{r}_n$ be the largest solution of the equation $\phi_n(r) = r$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ the following inequality holds uniformly for all $g \in \mathcal{G}$

$$\mathbb{E}_{\mathbf{z}}[g(\mathbf{z})] \leq \hat{\mathbb{E}}_{\widetilde{S}}[g] + 90(\hat{r}_n + r_0) + 4\sqrt{\hat{\mathbb{E}}_{\widetilde{S}}[g](\hat{r}_n + r_0)},$$

where $r_0 = b\big(\log(1/\delta) + 6\log\log n\big)/n$.

Proof of Theorem 4.13. For any $r > 0$, we define $\mathcal{F}_r$ as a subset of $\mathcal{F}$ with the empirical error less than or equal to r

$$\mathcal{F}_r = \Big\{f \in \mathcal{F} : \frac{1}{n}\sum_{j\in[n]} g_f(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_{j1}^-, \ldots, \mathbf{x}_{jk}^-) \leq r\Big\}.$$

Let

$$\Big\{\mathbf{r}^m = \big(r_{1,1}^m, r_{1,2}^m, \ldots, r_{1,k}^m, \ldots, r_{n,1}^m, r_{n,2}^m, \ldots, r_{n,k}^m\big) : m \in [N]\Big\}$$

be an $(\epsilon/(\sqrt{2r}G_s), \ell_\infty)$-cover of $\mathcal{H}_r := \{h_f \in \mathcal{H} : f \in \mathcal{F}_r\}$ w.r.t. $S_{\mathcal{H}}$. Then, by the definition of ℓ_∞-cover we know for any $f \in \mathcal{F}_r$ we can find $m \in [N]$ such that

$$\max_{j\in[n]} \max_{i\in[k]} \left|h_f(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_{ji}^-) - r_{j,i}^m\right| \leq \epsilon/(\sqrt{2r}G_s).$$

According to the self-bounding Lipschitz continuity of ℓ, we know

$$\begin{aligned}
&\frac{1}{n}\sum_{j\in[n]} \left|\ell\big(\{f(\mathbf{x}_j)^\top\big(f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-)\big)\}_{i=1}^k\big) - \ell\big(\{r_{j,i}^m\}_{i=1}^k\big)\right|^2 \\
&\leq \frac{G_s^2}{n}\sum_{j\in[n]} \max\Big\{\ell\big(\{f(\mathbf{x}_j)^\top\big(f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-)\big)\}_{i=1}^k\big), \ell\big(\{r_{j,i}^m\}_{i=1}^k\big)\Big\}\Big\|\big(f(\mathbf{x}_j)^\top\big(f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-)\big)\big)_{i=1}^k - \big(r_{j,i}^m\big)_{i=1}^k\Big\|_\infty^2 \\
&\leq \frac{G_s^2}{n}\sum_{j\in[n]} \Big(\ell\big(\{f(\mathbf{x}_j)^\top\big(f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-)\big)\}_{i=1}^k\big) + \ell\big(\{r_{j,i}^m\}_{i=1}^k\big)\Big)\Big\|\big(h_f(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_{ji}^-)\big)_{i=1}^k - \big(r_{j,i}^m\big)_{i=1}^k\Big\|_\infty^2 \\
&\leq 2G_s^2 r\epsilon^2/(2rG_s^2) = \epsilon^2,
\end{aligned}$$

where we have used the following inequalities due to the definition of $\mathcal{F}_r$

$$\frac{1}{n}\sum_{j\in[n]} \ell\big(\{f(\mathbf{x}_j)^\top\big(f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-)\big)\}_{i=1}^k\big) \leq r, \qquad \frac{1}{n}\sum_{j\in[n]} \ell\big(\{r_{j,i}^m\}_{i=1}^k\big) \leq r.$$

Therefore, we have

$$\mathcal{N}_2(\epsilon, \mathcal{G}_r, S) \leq \mathcal{N}_\infty(\epsilon/(\sqrt{2r}G_s), \mathcal{H}_r, S_{\mathcal{H}}),$$

where $\mathcal{G}_r = \{g_f \in \mathcal{G} : f \in \mathcal{F}_r\}$. Analyzing analogously to the proof of Theorem 4.8, we get (replacing G there by $\sqrt{2r}G_s$)

$$\mathfrak{R}_S(\mathcal{G}_r) \leq 24\sqrt{2r}G_s(R^2+1)n^{-\frac{1}{2}} + 48\sqrt{2r}G_s\sqrt{k}\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H})\bigg(1 + \log(4R^2n^{\frac{3}{2}}k)\Big\lceil \log_2 \frac{R^2\sqrt{n}}{12}\Big\rceil\bigg) := \psi_n(r).$$

Let $\hat{r}_n$ be the point satisfying $\hat{r}_n = \psi_n(\hat{r}_n)$:

$$\hat{r}_n = 24\sqrt{2\hat{r}_n}G_s(R^2+1)n^{-\frac{1}{2}} + 48\sqrt{2\hat{r}_n}G_s\sqrt{k}\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H})\bigg(1 + \log(4R^2n^{\frac{3}{2}}k)\Big\lceil \log_2 \frac{R^2\sqrt{n}}{12}\Big\rceil\bigg),$$

from which we get

$$\hat{r}_n = \widetilde{O}\Big(G_s^2R^4n^{-1} + G_s^2k\mathfrak{R}_{S_{\mathcal{H}},nk}^2(\mathcal{H})\Big).$$

We can apply Lemma D.1 to get the following inequality with probability at least $1-\delta$ uniformly for all $f \in \mathcal{F}$

$$L_{un}(f) = \hat{L}_{un}(f) + \widetilde{O}\Big(Bn^{-1} + G_s^2R^4n^{-1} + G_s^2k\mathfrak{R}_{S_{\mathcal{H}},nk}^2(\mathcal{H})\Big) + \widetilde{O}\Big(\sqrt{B}n^{-\frac{1}{2}} + G_sR^2n^{-\frac{1}{2}} + G_s\sqrt{k}\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H})\Big)\hat{L}_{un}^{\frac{1}{2}}(f).$$

We can apply Lemma 4.3 to control $\mathfrak{R}_{S_{\mathcal{H}},nk}(\mathcal{H})$ and derive the following bound

$$L_{un}(f) = \hat{L}_{un}(f) + \widetilde{O}\Big(Bn^{-1} + G_s^2R^4n^{-1} + G_s^2R^2n^{-2}k^{-1}\mathfrak{C}^2\Big) + \widetilde{O}\Big(\sqrt{B}n^{-\frac{1}{2}} + G_sR^2n^{-\frac{1}{2}} + G_sRn^{-1}k^{-\frac{1}{2}}\mathfrak{C}\Big)\hat{L}_{un}^{\frac{1}{2}}(f).$$

The proof is completed. □

E. Proof on Rademacher Complexities

We first prove Rademacher complexity bounds for feature spaces in Lemma 5.1, and then give lower bounds. Finally, we will apply it to prove Proposition 5.3 and Proposition 5.5.

Proof of Lemma 5.1. Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_d)^\top$ and $V_S = (\mathbf{v}(\mathbf{x}_1), \ldots, \mathbf{v}(\mathbf{x}_n))$. Then it is clear

$$UV_S = \begin{pmatrix} \mathbf{u}_1^\top\mathbf{v}(\mathbf{x}_1) & \cdots & \mathbf{u}_1^\top\mathbf{v}(\mathbf{x}_n) \\ \vdots & \vdots & \vdots \\ \mathbf{u}_d^\top\mathbf{v}(\mathbf{x}_1) & \cdots & \mathbf{u}_d^\top\mathbf{v}(\mathbf{x}_n) \end{pmatrix}.$$

Let

$$M_\epsilon = \begin{pmatrix} \epsilon_{1,1} & \cdots & \epsilon_{1,n} \\ \vdots & \vdots & \vdots \\ \epsilon_{d,1} & \cdots & \epsilon_{d,n} \end{pmatrix} \in \mathbb{R}^{d\times n}.$$

Then we have

$$\begin{aligned}\sum_{t\in[d]}\sum_{j\in[n]} \epsilon_{t,j}\mathbf{u}_t^\top \mathbf{v}(\mathbf{x}_j) &= \langle M_\epsilon, UV_S\rangle = \text{trace}(M_\epsilon^\top UV_S) = \text{trace}(UV_S M_\epsilon^\top) \\ &= \langle U^\top, V_S M_\epsilon^\top\rangle \le \|U^\top\|\|V_S M_\epsilon^\top\|_*,\end{aligned}$$

where trace denotes the trace of a matrix. Therefore, we have

$$\begin{aligned}\sup_{f\in\mathcal{F}}\sum_{t\in[d]}\sum_{j\in[n]}\epsilon_{t,j}f_t(\mathbf{x}_j) &= \sup_{U\in\mathcal{U},\mathbf{v}\in\mathcal{V}}\sum_{t\in[d]}\sum_{j\in[n]}\epsilon_{t,j}\mathbf{u}_t^\top\mathbf{v}(\mathbf{x}_j) \\ &\le \Lambda\sup_{\mathbf{v}\in\mathcal{V}}\|V_S M_\epsilon^\top\|_* = \Lambda\sup_{\mathbf{v}\in\mathcal{V}}\big\|\big(\sum_{j\in[n]}\epsilon_{1,j}\mathbf{v}(\mathbf{x}_j),\ldots,\sum_{j\in[n]}\epsilon_{d,j}\mathbf{v}(\mathbf{x}_j)\big)\big\|_*.\end{aligned}$$

The proof is completed. □

Proof of Lemma 5.2. Note

$$\begin{aligned}&\Big|\sum_{j\in[nk]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\tilde{\mathbf{x}}_j)+\epsilon_{j,t,2}f_t(\tilde{\mathbf{x}}_j^+)+\epsilon_{j,t,3}f_t(\tilde{\mathbf{x}}_j^-)\big)\Big| = \\ &\max\Big\{\sum_{j\in[nk]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\tilde{\mathbf{x}}_j)+\epsilon_{j,t,2}f_t(\tilde{\mathbf{x}}_j^+)+\epsilon_{j,t,3}f_t(\tilde{\mathbf{x}}_j^-)\big), -\sum_{j\in[nk]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\tilde{\mathbf{x}}_j)+\epsilon_{j,t,2}f_t(\tilde{\mathbf{x}}_j^+)+\epsilon_{j,t,3}f_t(\tilde{\mathbf{x}}_j^-)\big)\Big\}.\end{aligned}$$

According to the symmetry of $\mathcal{F}$ we know

$$\begin{aligned}\mathfrak{C} &= \max_{\{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_j^+,\tilde{\mathbf{x}}_j^-)\}_{j=1}^{nk}\subseteq S_\mathcal{H}}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nk}\times\{\pm1\}^d\times\{\pm1\}^3}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{j\in[nk]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\tilde{\mathbf{x}}_j)+\epsilon_{j,t,2}f_t(\tilde{\mathbf{x}}_j^+)+\epsilon_{j,t,3}f_t(\tilde{\mathbf{x}}_j^-)\big)\Big|\Big] \\ &\ge \sup_{f\in\mathcal{F}}\max_{\{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_j^+,\tilde{\mathbf{x}}_j^-)\}_{j=1}^{nk}\subseteq S_\mathcal{H}}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nk}\times\{\pm1\}^d\times\{\pm1\}^3}\Big[\Big|\sum_{j\in[nk]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\tilde{\mathbf{x}}_j)+\epsilon_{j,t,2}f_t(\tilde{\mathbf{x}}_j^+)+\epsilon_{j,t,3}f_t(\tilde{\mathbf{x}}_j^-)\big)\Big|\Big],\end{aligned}$$

where we have used the Jensen's inequality in the last step.

Since we take maximization over $\{(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_j^+,\tilde{\mathbf{x}}_j^-)\}_{j=1}^{nk}\subseteq S_\mathcal{H}$, we can choose $(\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_j^+,\tilde{\mathbf{x}}_j^-) = (\tilde{\mathbf{x}},\tilde{\mathbf{x}}^+,\tilde{\mathbf{x}}^-)$ for any $(\tilde{\mathbf{x}},\tilde{\mathbf{x}}^+,\tilde{\mathbf{x}}^-)\in S_\mathcal{H}$. Then we get

$$\begin{aligned}\mathfrak{C} &\ge \sup_{f\in\mathcal{F}}\max_{(\tilde{\mathbf{x}},\tilde{\mathbf{x}}^+,\tilde{\mathbf{x}}^-)\in S_\mathcal{H}}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nk}\times\{\pm1\}^d\times\{\pm1\}^3}\Big[\Big|\sum_{j\in[nk]}\sum_{t\in[d]}\big(\epsilon_{j,t,1}f_t(\tilde{\mathbf{x}})+\epsilon_{j,t,2}f_t(\tilde{\mathbf{x}}^+)+\epsilon_{j,t,3}f_t(\tilde{\mathbf{x}}^-)\big)\Big|\Big] \\ &\ge 2^{-\frac{1}{2}}\sup_{f\in\mathcal{F}}\max_{(\tilde{\mathbf{x}},\tilde{\mathbf{x}}^+,\tilde{\mathbf{x}}^-)\in S_\mathcal{H}}\Big(\sum_{j\in[nk]}\sum_{t\in[d]}\big(f_t^2(\tilde{\mathbf{x}})+f_t^2(\tilde{\mathbf{x}}^+)+f_t^2(\tilde{\mathbf{x}}^-)\big)\Big)^{\frac{1}{2}} \\ &= 2^{-\frac{1}{2}}\sup_{f\in\mathcal{F}}\max_{(\tilde{\mathbf{x}},\tilde{\mathbf{x}}^+,\tilde{\mathbf{x}}^-)\in S_\mathcal{H}}\Big(\sum_{j\in[nk]}\big(\|f(\tilde{\mathbf{x}})\|_2^2+\|f(\tilde{\mathbf{x}}^+)\|_2^2+\|f(\tilde{\mathbf{x}}^-)\|_2^2\big)\Big)^{\frac{1}{2}} \\ &= \sqrt{2^{-1}nk}\sup_{f\in\mathcal{F}}\max_{(\tilde{\mathbf{x}},\tilde{\mathbf{x}}^+,\tilde{\mathbf{x}}^-)\in S_\mathcal{H}}\Big(\|f(\tilde{\mathbf{x}})\|_2^2+\|f(\tilde{\mathbf{x}}^+)\|_2^2+\|f(\tilde{\mathbf{x}}^-)\|_2^2\Big)^{\frac{1}{2}},\end{aligned}$$

where we have used the following Khitchine-Kahane inequality (De la Pena & Giné, 2012)

$$\mathbb{E}_\epsilon\big|\sum_{i=1}^n\epsilon_i t_i\big| \ge 2^{-\frac{1}{2}}\big[\sum_{i=1}^n|t_i|^2\big]^{\frac{1}{2}},\quad \forall t_1,\ldots,t_n\in\mathbb{R}, \tag{E.1}$$

The proof is completed. □

Remark E.1. The analysis in the proof implies a lower bound for $\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}})$ for a symmetric $\widetilde{\mathcal{F}}$ and $\widetilde{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$

$$\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}}) \geq \frac{1}{\sqrt{2}n} \sup_{f \in \widetilde{\mathcal{F}}} \big\| \big(f(\mathbf{z}_1), \ldots, f(\mathbf{z}_n)\big) \big\|_2.$$

Indeed, by the symmetry of $\mathcal{F}$, the Jensen inequality and Eq. (E.1), we have

$$\begin{aligned}\mathfrak{R}_{\widetilde{S}}(\widetilde{\mathcal{F}}) &= \mathbb{E}_{\boldsymbol{\epsilon}}\big[\sup_{f \in \widetilde{\mathcal{F}}} \frac{1}{n} \sum_{i \in [n]} \epsilon_i f(\mathbf{z}_i)\big] = \mathbb{E}_{\boldsymbol{\epsilon}}\big[\sup_{f \in \widetilde{\mathcal{F}}} \frac{1}{n} \big| \sum_{i \in [n]} \epsilon_i f(\mathbf{z}_i)\big|\big] \\ &\geq \frac{1}{n} \sup_{f \in \widetilde{\mathcal{F}}} \mathbb{E}_{\boldsymbol{\epsilon}}\big[\big| \sum_{i \in [n]} \epsilon_i f(\mathbf{z}_i)\big|\big] \geq \frac{1}{\sqrt{2}n} \sup_{f \in \widetilde{\mathcal{F}}} \Big(\sum_{i \in [n]} f^2(\mathbf{z}_i)\Big)^{\frac{1}{2}}.\end{aligned}$$

E.1. Proof of Proposition 5.3

The following Khintchine-Kahane inequality (De la Pena & Giné, 2012; Lust-Piquard & Pisier, 1991) is very useful for us to estimate Rademacher complexities.

Lemma E.2. *Let $\epsilon_1, \ldots, \epsilon_n$ be a sequence of independent Rademacher variables.*

(a) Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with $\|\cdot\|$ being the associated norm. Then, for any $p \geq 1$ there holds

$$\big[\mathbb{E}_\epsilon \| \sum_{i=1}^n \epsilon_i \mathbf{v}_i \|^p\big]^{\frac{1}{p}} \leq \max(\sqrt{p-1}, 1) \big[\sum_{i=1}^n \|\mathbf{v}_i\|^2\big]^{\frac{1}{2}}. \tag{E.2}$$

(b) Let $X_1, \ldots, X_n$ be a set of matrices of the same dimension. For all $q \geq 2$,

$$\Big(\mathbb{E}_{\boldsymbol{\epsilon}} \big\| \sum_{i=1}^n \epsilon_i X_i \big\|_{S_q}^q\Big)^{\frac{1}{q}} \leq 2^{-\frac{1}{4}} \sqrt{\frac{q\pi}{e}} \max\Big\{ \big\| \big(\sum_{i=1}^n X_i^\top X_i\big)^{\frac{1}{2}} \big\|_{S_q}, \big\| \big(\sum_{i=1}^n X_i X_i^\top\big)^{\frac{1}{2}} \big\|_{S_q} \Big\}. \tag{E.3}$$

Proof of Proposition G.1. Let $q \geq p$. It is clear $q^* \leq p^*$. The dual norm of $\|\cdot\|_{2,p}$ is $\|\cdot\|_{2,p^*}$. Therefore, according to Lemma 5.1 and $\|\cdot\|_{p^*} \leq \|\cdot\|_{q^*}$ we know

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f \in \mathcal{F}} \sum_{t \in [d]} \sum_{j \in [n]} \epsilon_{t,j} f_t(\mathbf{x}_j) &\leq \Lambda \mathbb{E}_{\boldsymbol{\epsilon}} \Big(\sum_{t \in [d]} \big\| \sum_{j \in [n]} \epsilon_{t,j} \mathbf{x}_j \big\|_2^{p^*} \Big)^{1/p^*} \leq \Lambda \mathbb{E}_{\boldsymbol{\epsilon}} \Big(\sum_{t \in [d]} \big\| \sum_{j \in [n]} \epsilon_{t,j} \mathbf{x}_j \big\|_2^{q^*} \Big)^{1/q^*} \\ &\leq \Lambda \Big(\mathbb{E}_{\boldsymbol{\epsilon}} \Big[\sum_{t \in [d]} \big\| \sum_{j \in [n]} \epsilon_{t,j} \mathbf{x}_j \big\|_2^{q^*} \Big] \Big)^{1/q^*} = \Lambda \Big(d \mathbb{E}_{\boldsymbol{\epsilon}} \big\| \sum_{j \in [n]} \epsilon_j \mathbf{x}_j \big\|_2^{q^*} \Big)^{1/q^*},\end{aligned}$$

where we have used Jense's inequality and the concavity of $x \mapsto x^{1/q^*}$. By Lemma E.2, we know

$$\mathbb{E}_{\boldsymbol{\epsilon}} \big\| \sum_{j \in [n]} \epsilon_j \mathbf{x}_j \big\|_2^{q^*} \leq \max(\sqrt{q^*-1}, 1)^{q^*} \Big(\sum_{j \in [n]} \|\mathbf{x}_j\|_2^2 \Big)^{\frac{q^*}{2}}.$$

It then follows that

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f \in \mathcal{F}} \sum_{t \in [d]} \sum_{j \in [n]} \epsilon_{t,j} f_t(\mathbf{x}_j) \leq \Lambda d^{1/q^*} \max(\sqrt{q^*-1}, 1) \Big(\sum_{j \in [n]} \|\mathbf{x}_j\|_2^2 \Big)^{\frac{1}{2}}.$$

Note the above inequality holds for any $q \geq p$. This proves Part (a).

We now prove Part (b). Since the dual norm of $\|\cdot\|_{S_p}$ is $\|\cdot\|_{S_{p^*}}$, by Lemma 5.1 we know

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f \in \mathcal{F}} \sum_{t \in [d]} \sum_{j \in [n]} \epsilon_{t,j} f_t(\mathbf{x}_j) \leq \Lambda \mathbb{E}_{\boldsymbol{\epsilon}} \big\| \big(\sum_{j \in [n]} \epsilon_{1,j} \mathbf{x}_j, \ldots, \sum_{j \in [n]} \epsilon_{d,j} \mathbf{x}_j \big) \big\|_{S_{p^*}}.$$

For any $t \in [d]$ and $j \in [n]$, define

$$\widetilde{X}_{t,j} = \begin{pmatrix} 0 & \cdots & 0 & \mathbf{x}_j & 0 & \cdots & 0 \end{pmatrix},$$

i.e., the t-th column of $\widetilde{X}_{t,j} = \mathbf{x}_j$, and other columns are zero vectors. This implies that

$$\big(\sum_{j\in[n]} \epsilon_{1,j}\mathbf{x}_j, \ldots, \sum_{j\in[n]} \epsilon_{d,j}\mathbf{x}_j\big) = \sum_{t\in[d]}\sum_{j\in[n]} \epsilon_{t,j}\widetilde{X}_{t,j}.$$

It is clear that $\widetilde{X}_{t,j}\widetilde{X}_{t,j}^\top = \mathbf{x}_j\mathbf{x}_j^\top$ and

$$\widetilde{X}_{t,j}^\top\widetilde{X}_{t,j} = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & \cdots & \cdots & 0 \\ \vdots & 0 & \mathbf{x}_j^\top\mathbf{x}_j & 0 \\ 0 & \cdots & \cdots & \ldots. \end{pmatrix} = \mathbf{x}_j^\top\mathbf{x}_j \text{diag}(\underbrace{0,\ldots,0}_{t-1}, 1, 0\ldots, 0),$$

where $\text{diag}(a_1, \ldots, a_n)$ denotes the diagonal matrix with elements $a_1, \ldots, a_n$. Therefore, we have

$$\sum_{t\in[d]}\sum_{j\in[n]} \widetilde{X}_{t,j}\widetilde{X}_{t,j}^\top = d\sum_{j\in[n]} \mathbf{x}_j\mathbf{x}_j^\top$$

and

$$\sum_{t\in[d]}\sum_{j\in[n]} \widetilde{X}_{t,j}^\top\widetilde{X}_{t,j} = \big(\sum_{j\in[n]} \mathbf{x}_j^\top\mathbf{x}_j\big)\mathbb{I}_{d\times d},$$

where $\mathbb{I}_{d\times d}$ denotes the identity matrix in $\mathbb{R}^{d\times d}$. Therefore, we can apply Lemma E.2 to show that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]}\sum_{j\in[n]} \epsilon_{t,j} f_t(\mathbf{x}_j) &\le \Lambda\Big(\mathbb{E}_{\boldsymbol{\epsilon}}\big\|\big(\sum_{j\in[n]} \epsilon_{1,j}\mathbf{x}_j, \ldots, \sum_{j\in[n]} \epsilon_{d,j}\mathbf{x}_j\big)\big\|_{S_q^*}^{q^*}\Big)^{1/q^*} \\ &\le \Lambda 2^{-\frac{1}{4}}\sqrt{\frac{q^*\pi}{e}} \max\Big\{\Big\|\Big(d\sum_{j\in[n]} \mathbf{x}_j\mathbf{x}_j^\top\Big)^{\frac{1}{2}}\Big\|_{S_{q^*}}, d^{1/q^*}\big(\sum_{j\in[d]} \|\mathbf{x}_j\|_2^2\big)^{\frac{1}{2}}\Big\}.\end{aligned}$$

The proof is completed. □

E.2. Proof of Proposition 5.5

For convenience we introduce the following sequence of function spaces

$$\mathcal{V}_k = \Big\{\mathbf{x}\mapsto \sigma_k\big(V_k\sigma\big(V_{k-1}\cdots\sigma(V_1\mathbf{x})\big)\big) : \|V_j\|_F \le B_j\Big\}, \quad k\in[L].$$

To prove Proposition 5.5, we need to introduce several lemmas. The following lemma shows how the supremum over a matrix can be transferred to a supremum over a vector. It is an extension of Lemma 1 in Golowich et al. (2018) from $d = 1$ to $d \in \mathbb{N}$, and can be proved exactly by the arguments in Golowich et al. (2018).

Lemma E.3. *Let $\sigma : \mathbb{R}\mapsto\mathbb{R}$ be a 1-Lipschitz continuous, positive-homogeneous activation function which is applied elementwise. Then for any vector-valued function class $\widetilde{\mathcal{F}}$*

$$\sup_{\tilde{f}\in\widetilde{\mathcal{F}}, V\in\mathbb{R}^{h\times h'}:\|V\|_F\le B} \sum_{t\in[d]} \Big\|\sum_{j\in[n]} \epsilon_{t,j}\sigma(V\tilde{f}(\mathbf{x}_j))\Big\|_2^2 \le B^2 \sup_{\tilde{f}\in\widetilde{\mathcal{F}}, \tilde{\mathbf{v}}\in\mathbb{R}^{h'}:\|\tilde{\mathbf{v}}\|_2\le 1} \sup_{\|\tilde{\mathbf{v}}\|_2\le 1} \sum_{t\in[d]} \Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j))\Big|^2.$$

Proof. Let $\mathbf{v}_1^\top, \ldots, \mathbf{v}_h^\top$ be rows of matrix V, i.e., $V^\top = \big(\mathbf{v}_1, \ldots, \mathbf{v}_h\big)$. Then by the positive-homogeneous property of activation function we have

$$\begin{aligned}\sum_{t\in[d]} \Big\|\sum_{j\in[n]} \epsilon_{t,j}\sigma(V\tilde{f}(x_i))\Big\|_2^2 &= \sum_{t\in[d]} \left\|\begin{pmatrix} \sum_{j\in[n]} \epsilon_{t,j}\sigma(\mathbf{v}_1^\top\tilde{f}(\mathbf{x}_j)) \\ \vdots \\ \sum_{j\in[n]} \epsilon_{t,j}\sigma(\mathbf{v}_h^\top\tilde{f}(\mathbf{x}_j)) \end{pmatrix}\right\|_2^2 = \sum_{t\in[d]}\sum_{r\in[h]} \left(\sum_{j\in[n]} \epsilon_{t,j}\sigma(\mathbf{v}_r^\top\tilde{f}(\mathbf{x}_j))\right)^2 \\ &= \sum_{r\in[h]} \|\mathbf{v}_r\|_2^2 \sum_{t\in[d]} \left(\sum_{j\in[n]} \epsilon_{t,j}\sigma\Big(\frac{\mathbf{v}_r^\top}{\|\mathbf{v}_r\|_2}\tilde{f}(\mathbf{x}_j)\Big)\right)^2 \\ &\le \Big(\sum_{r\in[h]} \|\mathbf{v}_r\|_2^2\Big) \max_{r\in[h]} \sum_{t\in[d]} \Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma\Big(\frac{\mathbf{v}_r^\top}{\|\mathbf{v}_r\|_2}\tilde{f}(\mathbf{x}_j)\Big)\Big|^2 \\ &\le B^2 \sup_{\|\tilde{\mathbf{v}}\|_2\le 1} \sum_{t\in[d]} \Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j))\Big|^2.\end{aligned}$$

The proof is completed. □

The following lemma gives a general contraction lemma for Rademacher complexities. It allows us to remove a nonlinear function ψ, which is very useful for us to handle the activation function in DNNs.

Lemma E.4 (Contraction Lemma, Thm 11.6 in Boucheron et al. (2013)). *Let $\tilde{\tau}: \mathbb{R}_+ \mapsto \mathbb{R}_+$ be convex and nondecreasing. Suppose $\psi: \mathbb{R} \mapsto \mathbb{R}$ is contractive in the sense $|\psi(t) - \psi(\tilde{t})| \leq |t - \tilde{t}|$ and $\psi(0) = 0$. Then the following inequality holds for any $\widetilde{\mathcal{F}}$*

$$\mathbb{E}_{\boldsymbol{\epsilon}} \tilde{\tau}\bigg(\sup_{f \in \widetilde{\mathcal{F}}} \sum_{i=1}^n \epsilon_i \psi\big(f(x_i)\big) \bigg) \leq \mathbb{E}_{\boldsymbol{\epsilon}} \tilde{\tau}\bigg(\sup_{f \in \widetilde{\mathcal{F}}} \sum_{i=1}^n \epsilon_i f(x_i) \bigg).$$

The following lemma gives bounds of MGFs for a random variable $Z = \sum_{1 \leq i < j \leq n} \epsilon_i \epsilon_j a_{ij}$, which is called a Rademacher chaos variable (De la Pena & Giné, 2012; Ying & Campbell, 2010).

Lemma E.5 (page 167 in De la Pena & Giné (2012)). *Let $\epsilon_i, i \in [n]$ be independent Rademacher variables. Let $a_{i,j} \in \mathbb{R}, i, j \in [n]$. Then for $Z = \sum_{1 \leq i < j \leq n} \epsilon_i \epsilon_j a_{ij}$ we have*

$$\mathbb{E}_{\boldsymbol{\epsilon}} \exp\Big(|Z|/(4es) \Big) \leq 2, \quad \textit{where } s^2 := \sum_{1 \leq i < j \leq n} a_{i,j}^2.$$

Proof of Proposition 5.5. The dual norm of $\|\cdot\|_F$ is $\|\cdot\|_F$. Therefore, according to Lemma 5.1 we know

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f \in \mathcal{F}} \sum_{t \in [d]} \sum_{j \in [n]} \epsilon_{t,j} f_t(\mathbf{x}_j) &\leq \Lambda \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\mathbf{v} \in \mathcal{V}} \Big(\sum_{t \in [d]} \Big\| \sum_{j \in [n]} \epsilon_{t,j} \mathbf{v}(\mathbf{x}_j) \Big\|_2^2 \Big)^{1/2} \\
&= \Lambda \mathbb{E}_{\boldsymbol{\epsilon}} \bigg(\sup_{\tilde{f} \in \mathcal{V}_{L-1}, V: \|V\|_F \leq B_L} \sum_{t \in [d]} \Big\| \sum_{j \in [n]} \epsilon_{t,j} \sigma(V \tilde{f}(\mathbf{x}_j)) \Big\|_2^2 \bigg)^{\frac{1}{2}} \\
&\leq \Lambda B_L \mathbb{E}_{\boldsymbol{\epsilon}} \bigg(\sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \sum_{t \in [d]} \Big| \sum_{j \in [n]} \epsilon_{t,j} \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \bigg)^{\frac{1}{2}},
\end{aligned}$$

where we have used Lemma E.3 in the second inequality. Let $\lambda \geq 0$ and $\tau(x) = \exp(\lambda x^2)$. It is clear that τ is convex and increasing in the interval $[0, \infty)$. It then follows from the Jensen's inequality that

$$\begin{aligned}
\exp\bigg(\lambda \Big(\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f \in \mathcal{F}} \sum_{t \in [d]} \sum_{j \in [n]} \epsilon_{t,j} f_t(\mathbf{x}_j) \Big)^2 \bigg) &\leq \exp\bigg(\lambda \Big(\Lambda B_L \mathbb{E}_{\boldsymbol{\epsilon}} \Big(\sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \sum_{t \in [d]} \Big| \sum_{j \in [n]} \epsilon_{t,j} \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \Big)^{\frac{1}{2}} \Big)^2 \bigg) \\
&\leq \mathbb{E}_{\boldsymbol{\epsilon}} \exp\bigg(\lambda \Big(\Lambda B_L \Big(\sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \sum_{t \in [d]} \Big| \sum_{j \in [n]} \epsilon_{t,j} \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \Big)^{\frac{1}{2}} \Big)^2 \bigg) \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \exp\bigg(\lambda \Lambda^2 B_L^2 \sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \sum_{t \in [d]} \Big| \sum_{j \in [n]} \epsilon_{t,j} \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \bigg) \\
&\leq \mathbb{E}_{\boldsymbol{\epsilon}} \exp\bigg(\lambda \Lambda^2 B_L^2 \sum_{t \in [d]} \sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \Big| \sum_{j \in [n]} \epsilon_{t,j} \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \bigg) \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \prod_{t \in [d]} \exp\bigg(\lambda \Lambda^2 B_L^2 \sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \Big| \sum_{j \in [n]} \epsilon_{t,j} \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \bigg) \\
&= \prod_{t \in [d]} \mathbb{E}_{\boldsymbol{\epsilon}_t} \exp\bigg(\lambda \Lambda^2 B_L^2 \sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \Big| \sum_{j \in [n]} \epsilon_{t,j} \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \bigg) \\
&= \mathbb{E}_{\boldsymbol{\epsilon} \sim \{\pm 1\}^n} \exp\bigg(d \lambda \Lambda^2 B_L^2 \sup_{\tilde{f} \in \mathcal{V}_{L-1}, \tilde{\mathbf{v}}: \|\tilde{\mathbf{v}}\|_2 \leq 1} \Big| \sum_{j \in [n]} \epsilon_j \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j)) \Big|^2 \bigg),
\end{aligned}$$

where we have used the independency between $\epsilon_t = (\epsilon_{t,j})_{j\in[n]}, t \in [d]$. Let $\tilde{\tau} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be defined as $\tilde{\tau}(x) = \exp(d\lambda\Lambda^2 B_L^2 x^2)$. Then we have

$$
\begin{aligned}
&\exp\bigg(\lambda\Big(\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]}\sum_{j\in[n]} \epsilon_{t,j} f_t(\mathbf{x}_j)\Big)^2\bigg) \leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1},\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\leq 1}\Big|\sum_{j\in[n]}\epsilon_j\sigma(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j))\Big|\bigg)\\
&\leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1},\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\leq 1}\sum_{j\in[n]}\epsilon_j\sigma(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j))\bigg) + \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1},\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\leq 1} -\sum_{j\in[n]}\epsilon_j\sigma(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j))\bigg)\\
&= 2\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1},\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\leq 1}\sum_{j\in[n]}\epsilon_j\sigma(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j))\bigg) \leq 2\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1},\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\leq 1}\sum_{j\in[n]}\epsilon_j\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j)\bigg)\\
&= \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1},\tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\leq 1}\tilde{\mathbf{v}}^\top\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\bigg) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1}}\Big\|\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\Big\|_2\bigg). \qquad \text{(E.4)}
\end{aligned}
$$

where we have used Lemma E.4 and the contraction property of σ in the last inequality.

According to Lemma E.3, we know

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1}}\Big\|\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\Big\|_2\bigg) &= \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\|V_{L-1}\|_F\leq B_{L-1},\tilde{f}\in\mathcal{V}_{L-2}}\Big\|\sum_{j\in[n]}\epsilon_j\sigma\big(V_{L-1}\tilde{f}(\mathbf{x}_j)\big)\Big\|_2\bigg)\\
&\leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\sup_{\|\tilde{\mathbf{v}}\|_2\leq 1,\tilde{f}\in\mathcal{V}_{L-2}}\Big|\sum_{j\in[n]}\epsilon_j\sigma\big(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j)\big)\Big|\bigg).
\end{aligned}
$$

It then follows that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1}}\Big\|\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\Big\|_2\bigg)\\
&\leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\sup_{\|\tilde{\mathbf{v}}\|_2\leq 1,\tilde{f}\in\mathcal{V}_{L-2}}\sum_{j\in[n]}\epsilon_j\sigma\big(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j)\big)\bigg) + \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\sup_{\|\tilde{\mathbf{v}}\|_2\leq 1,\tilde{f}\in\mathcal{V}_{L-2}} -\sum_{j\in[n]}\epsilon_j\sigma\big(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j)\big)\bigg)\\
&= 2\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\sup_{\|\tilde{\mathbf{v}}\|_2\leq 1,\tilde{f}\in\mathcal{V}_{L-2}}\sum_{j\in[n]}\epsilon_j\sigma\big(\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j)\big)\bigg) \leq 2\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\sup_{\|\tilde{\mathbf{v}}\|_2\leq 1,\tilde{f}\in\mathcal{V}_{L-2}}\sum_{j\in[n]}\epsilon_j\tilde{\mathbf{v}}^\top\tilde{f}(\mathbf{x}_j)\bigg)\\
&= 2\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\sup_{\|\tilde{\mathbf{v}}\|_2\leq 1,\tilde{f}\in\mathcal{V}_{L-2}}\tilde{\mathbf{v}}^\top\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\bigg) \leq 2\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\sup_{\tilde{f}\in\mathcal{V}_{L-2}}\Big\|\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\Big\|_2\bigg).
\end{aligned}
$$

We can apply the above inequality recursively and derive

$$
\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1}}\Big\|\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\Big\|_2\bigg) \leq 2^{L-1}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\cdots B_1\Big\|\sum_{j\in[n]}\epsilon_j\mathbf{x}_j\Big\|_2\bigg).
$$

Furthermore, by Eq. (E.4) we know

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{t\in[d]}\sum_{j\in[n]}\epsilon_{t,j}f_t(\mathbf{x}_j) &= \tau^{-1}\tau\bigg(\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{t\in[d]}\sum_{j\in[n]}\epsilon_{t,j}f_t(\mathbf{x}_j)\bigg)\\
&\leq \tau^{-1}\bigg(\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(\sup_{\tilde{f}\in\mathcal{V}_{L-1}}\Big\|\sum_{j\in[n]}\epsilon_j\tilde{f}(\mathbf{x}_j)\Big\|_2\bigg)\bigg)\\
&\leq \tau^{-1}\bigg(2^{L-1}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\bigg(B_{L-1}\cdots B_1\Big\|\sum_{j\in[n]}\epsilon_j\mathbf{x}_j\Big\|_2\bigg)\bigg)\\
&= \tau^{-1}\bigg(2^{L-1}\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\exp\Big(d\lambda\Lambda^2B_L^2B_{L-1}^2\cdots B_1^2\Big\|\sum_{j\in[n]}\epsilon_j\mathbf{x}_j\Big\|_2^2\Big)\bigg),
\end{aligned}
$$

where the last identity follows from the definition of $\tilde{\tau}$. Let $\lambda_0 = d\lambda\Lambda^2 B_L^2 B_{L-1}^2 \cdots B_1^2$. Then

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \exp\Big(\lambda_0 \Big\| \sum_{j\in[n]} \epsilon_j \mathbf{x}_j \Big\|_2^2\Big) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \exp\Big(\lambda_0 \sum_{j\in[n]} \|\mathbf{x}_j\|_2^2 + 2\lambda_0 \sum_{1\le i<j\le n} \epsilon_i\epsilon_j \mathbf{x}_i^\top \mathbf{x}_j\Big)$$
$$\le \exp\Big(\lambda_0 \sum_{j\in[n]} \|\mathbf{x}_j\|_2^2\Big) \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \exp\Big(2\lambda_0 \sum_{1\le i<j\le n} \epsilon_i\epsilon_j \mathbf{x}_i^\top \mathbf{x}_j\Big).$$

We choose $\lambda = \frac{1}{8esd\Lambda^2 B_L^2 B_{L-1}^2 \cdots B_1^2}$, where $s = \big(\sum_{1\le i<j\le n} (\mathbf{x}_i^\top \mathbf{x}_j)^2\big)^{\frac{1}{2}}$. Then it is clear $\lambda_0 = \frac{1}{8es}$. We can apply Lemma E.5 to derive that

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \exp\Big(2\lambda_0 \sum_{1\le i<j\le n} \epsilon_i\epsilon_j \mathbf{x}_i^\top \mathbf{x}_j\Big) \le 2$$

and therefore

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \exp\Big(\lambda_0 \Big\| \sum_{j\in[n]} \epsilon_j \mathbf{x}_j \Big\|_2^2\Big) \le 2\exp\Big(\lambda_0 \sum_{j\in[n]} \|\mathbf{x}_j\|_2^2\Big).$$

We know $\tau^{-1}(x) = \sqrt{\lambda^{-1}\log x}$. It then follows that

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]} \sum_{j\in[n]} \epsilon_{t,j} f_t(\mathbf{x}_j) \le \Big(\lambda^{-1}(L-1)\log 2 + \lambda^{-1}\log \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \exp\Big(\lambda_0 \Big\| \sum_{j\in[n]} \epsilon_j \mathbf{x}_j \Big\|_2^2\Big)\Big)^{\frac{1}{2}}$$
$$\le \Big(\lambda^{-1}(L-1)\log 2 + \lambda^{-1}\log\Big(2\exp\Big(\lambda_0 \sum_{j\in[n]} \|\mathbf{x}_j\|_2^2\Big)\Big)\Big)^{\frac{1}{2}}$$
$$= \Big(\lambda^{-1} L\log 2 + \lambda^{-1}\lambda_0 \sum_{j\in[n]} \|\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}$$
$$= \Big(8esd\Lambda^2 B_L^2 B_{L-1}^2 \cdots B_1^2 L\log 2 + d\Lambda^2 B_L^2 B_{L-1}^2 \cdots B_1^2 \sum_{j\in[n]} \|\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}$$
$$= \sqrt{d}\Lambda B_L B_{L-1} \cdots B_1 \Big(8esL\log 2 + \sum_{j\in[n]} \|\mathbf{x}_j\|_2^2\Big)^{\frac{1}{2}}.$$

The proof is completed by noting $8e(\log 2) \le 16$. □

Remark E.6. Our proof of Proposition 5.5 is motivated by the arguments in Golowich et al. (2018), which studies Rademacher complexity bounds for DNNs with $d = 1$. Our analysis requires to introduce techniques to handle the difficulty of considering d features simultaneously. Indeed, we control the Rademacher complexity for learning with d features by

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]} \sum_{j\in[n]} \epsilon_{t,j} f_t(\mathbf{x}_j) \le \mathbb{E}_{\boldsymbol{\epsilon}}\Big(\sup_{\tilde{f}\in\mathcal{V}_{L-1}, \tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\le 1} \sum_{t\in[d]} \Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j))\Big|^2\Big)^{\frac{1}{2}}.$$

If $d = 1$, this becomes

$$\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \sum_{t\in[d]} \sum_{j\in[n]} \epsilon_{t,j} f_t(\mathbf{x}_j) \le \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\tilde{f}\in\mathcal{V}_{L-1}, \tilde{\mathbf{v}}:\|\tilde{\mathbf{v}}\|_2\le 1} \Big|\sum_{j\in[n]} \epsilon_j \sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j))\Big|,$$

and the arguments in Golowich et al. (2018) apply. There are two difficulties in applying the arguments in Golowich et al. (2018) to handle general $d \in \mathbb{N}$. First, the term $\sum_{t\in[d]} \Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j))\Big|$ cannot be written as a Rademacher complexity due to the summation over $t \in [d]$. Second, there is a square function of the term $\Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j))\Big|$. To handle this difficulty, we introduce the function $\tau(x) = \exp(\lambda x^2)$ instead of the function $\tau(x) = \exp(\lambda x)$ in Golowich et al. (2018). To this aim, we need to handle the MGF of a Rademacher chaos variable $\sum_{1\le i<j\le j} \epsilon_i\epsilon_j(\mathbf{x}_i^\top \mathbf{x}_j)^2$, which is not a sub-Gaussian variable. As a comparison, the analysis in Golowich et al. (2018) considers the MGF for a sub-Gaussian variable. One can also use the following inequality

$$\Big(\sum_{t\in[d]} \Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j))\Big|^2\Big)^{\frac{1}{2}} \le \sum_{t\in[d]} \Big|\sum_{j\in[n]} \epsilon_{t,j}\sigma(\tilde{\mathbf{v}}^\top \tilde{f}(\mathbf{x}_j))\Big|,$$

the latter of which can then be further controlled by the arguments in Golowich et al. (2018). This, however, incurs a bound with a linear dependency on d. As a comparison, our analysis gives a bound with a square-root dependency on d.

F. A General Vector-contraction Inequality for Rademacher Complexities

In this section, we provide a general vector-contraction inequality for Rademacher complexities, which recovers Lemma B.1 with $\tau(a) = a$. The lemma is motivated from Lemma E.4 by considering a general convex and nondecreasing τ.

Theorem F.1. *Let $\mathcal{F}$ be a class of bounded functions $f : \mathcal{Z} \mapsto \mathbb{R}^d$ which contains the zero function. Let $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous, non-decreasing and convex function. Assume $\tilde{g}_1, \ldots, \tilde{g}_n : \mathbb{R}^d \to \mathbb{R}$ are G-Lipschitz continuous w.r.t. $\|\cdot\|_2$ and satisfy $\tilde{g}_i(\mathbf{0}) = 0$. Then*

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tau\Big(\sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_i\tilde{g}_i(f(\mathbf{x}_i))\Big) \leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}}\tau\Big(G\sqrt{2}\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sum_{j=1}^d \epsilon_{i,j}f_j(\mathbf{x}_i)\Big). \tag{F.1}$$

The following lemma is due to (Maurer, 2016). We provide here the proof for completeness.

Lemma F.2. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{Z} \mapsto \mathbb{R}^d$ and g be any functional defined on $\mathcal{F}$. Assume that $\tilde{g}_1, \ldots, \tilde{g}_n : \mathbb{R}^d \to \mathbb{R}$ are G-Lipschitz continuous w.r.t. $\|\cdot\|_2$. Then,*

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\sup_{f\in\mathcal{F}}\big[g(f) + \sum_{i=1}^n \epsilon_i\tilde{g}_i(f(\mathbf{x}_i))\big] \leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}}\sup_{f\in\mathcal{F}}\Big[g(f) + G\sqrt{2}\sum_{i=1}^n\sum_{j=1}^d \epsilon_{i,j}f_j(\mathbf{x}_i)\Big]. \tag{F.2}$$

Proof. We prove this result by induction. According to the symmetry between f and $\tilde{f}$, we derive

$$\begin{aligned}
&\mathbb{E}_{\epsilon_n}\sup_{f\in\mathcal{F}}\big[g(f) + \sum_{i=1}^n \epsilon_i\tilde{g}_i(f(\mathbf{x}_i))\big] \\
&= \frac{1}{2}\sup_{f,\tilde{f}\in\mathcal{F}}\Big[g(f) + g(\tilde{f}) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(f(\mathbf{x}_i)) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(\tilde{f}(\mathbf{x}_i)) + \tilde{g}_n(f(\mathbf{x}_n)) - \tilde{g}_n(\tilde{f}(\mathbf{x}_n))\Big] \\
&= \frac{1}{2}\sup_{f,\tilde{f}\in\mathcal{F}}\Big[g(f) + g(\tilde{f}) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(f(\mathbf{x}_i)) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(\tilde{f}(\mathbf{x}_i)) + \big|\tilde{g}_n(f(\mathbf{x}_n)) - \tilde{g}_n(\tilde{f}(\mathbf{x}_n))\big|\Big],
\end{aligned} \tag{F.3}$$

According to the Lipschitz property and Eq. (E.1), we derive

$$\big|\tilde{g}_n(f(\mathbf{x}_n)) - \tilde{g}_n(\tilde{f}(\mathbf{x}_n))\big| \leq G\big\|f(\mathbf{x}_n) - \tilde{f}(\mathbf{x}_n)\big\|_2 \leq G\sqrt{2}\mathbb{E}_{\epsilon_{n,1},\ldots,\epsilon_{n,j}}\big|\sum_{j=1}^d \epsilon_{n,j}\big[f_j(\mathbf{x}_n) - \tilde{f}_j(\mathbf{x}_n)\big]\big|.$$

Plugging the above inequality back into (F.3) and using the Jensen's inequality, we get

$$\begin{aligned}
&\mathbb{E}_{\epsilon_n}\sup_{f\in\mathcal{F}}\big[g(f) + \sum_{i=1}^n \epsilon_i\tilde{g}_i(f(\mathbf{x}_i))\big] \\
&\leq \frac{1}{2}\mathbb{E}_{\epsilon_{n,1},\ldots,\epsilon_{n,j}}\sup_{f,\tilde{f}\in\mathcal{F}}\Big[g(f) + g(\tilde{f}) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(f(\mathbf{x}_i)) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(\tilde{f}(\mathbf{x}_i)) + G\sqrt{2}\big|\sum_{j=1}^d \epsilon_{n,j}\big[f_j(\mathbf{x}_n) - \tilde{f}_j(\mathbf{x}_n)\big]\big|\Big] \\
&= \frac{1}{2}\mathbb{E}_{\epsilon_{n,1},\ldots,\epsilon_{n,j}}\sup_{f,\tilde{f}\in\mathcal{F}}\Big[g(f) + g(\tilde{f}) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(f(\mathbf{x}_i)) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(\tilde{f}(\mathbf{x}_i)) + G\sqrt{2}\sum_{j=1}^d \epsilon_{n,j}\big[f_j(\mathbf{x}_n) - \tilde{f}_j(\mathbf{x}_n)\big]\Big] \\
&= \mathbb{E}_{\epsilon_{n,1},\ldots,\epsilon_{n,j}}\sup_{f\in\mathcal{F}}\Big[g(f) + \sum_{i=1}^{n-1} \epsilon_i\tilde{g}_i(f(\mathbf{x}_i)) + G\sqrt{2}\sum_{j=1}^d \epsilon_{n,j}f_j(\mathbf{x}_n)\Big],
\end{aligned}$$

where we have used the symmetry in the second step.

The stated result can be derived by continuing the above deduction with expectation over ϵ_{n-1}, ϵ_{n-2} and so on. □

To prove Theorem F.1, we introduce the following lemmas on the approximation of a continuous, non-decreasing and convex function. Let $a_+ = \max\{a, 0\}$.

Lemma F.3. *Let $f : [a, b] \to \mathbb{R}_+$ be a continuous, non-decreasing and convex function and $m \geq 2$. Let $a = x_1 < \cdots < x_m = b$. Then the function $\tilde{g} : [a, b] \to \mathbb{R}$ defined by*

$$\tilde{g}(x) = f(x_k) + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}(x - x_k), \quad \text{if } x \in [x_k, x_{k+1}]$$

belongs to the set

$$H_{[a,b]} := \Big\{ c_0 + \sum_{i=1}^{m} c_i (x - t_i)_+ : c_i \geq 0, i \in [n], t_i \in \mathbb{R}, m \in \mathbb{N}, x \in [a, b] \Big\}. \tag{F.4}$$

Proof. Define

$$\bar{f}(x) = f(x_1) + \sum_{i=1}^{m-1} \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \big[(x - x_i)_+ - (x - x_{i+1})_+ \big].$$

We first show that $\bar{f}(x) = \tilde{g}(x)$ for all $x \in [a, b]$. Suppose that $x \in [x_k, x_{k+1})$. Then, it is clear that

$$\begin{aligned}
\bar{f}(x) &= f(x_1) + \sum_{i=1}^{k-1} \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \big[(x - x_i)_+ - (x - x_{i+1})_+ \big] + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} \big[(x - x_k)_+ - (x - x_{k+1})_+ \big] \\
&\quad + \sum_{i=k+1}^{m-1} \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \big[(x - x_i)_+ - (x - x_{i+1})_+ \big] \\
&= f(x_1) + \sum_{i=1}^{k-1} \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \big[(x - x_i) - (x - x_{i+1}) \big] + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} \big[(x - x_k) - 0 \big] \\
&= f(x_k) + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}(x - x_k) = \tilde{g}(x).
\end{aligned}$$

We now show that $\bar{f}(x)$ belongs to the set $H_{[a,b]}$. Indeed, it follows from $(x - x_m)_+ = 0$ for all $x \leq x_m = b$ that

$$\begin{aligned}
\bar{f}(x) &= f(x_1) + \sum_{i=1}^{m-1} \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}(x - x_i)_+ - \sum_{i=2}^{m} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}(x - x_i)_+ \\
&= f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1)_+ + \sum_{i=2}^{m-1} \Big[\frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} - \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \Big](x - x_i)_+.
\end{aligned}$$

Therefore, $\bar{f}(x)$ can be written as $\bar{f}(x) = c_0 + \sum_{i=1}^{m-1} c_i (x - t_i)_+$ with $t_i = x_i, c_0 = f(x_1), c_1 = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$ and $c_i = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} - \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}, i = 2, \ldots, m - 1$. The terms $c_1, \ldots, c_{m-1}$ are all non-negative since f is non-decreasing and convex. The proof is completed. □

Lemma F.4. *If $f : [a, b] \to \mathbb{R}_+$ is continuous, non-decreasing and convex, then f belongs to the closure of $H_{[a,b]}$ defined in Eq. (F.4).*

Proof. Let $m \in \mathbb{N}$. We can find $a = x_1^{(m)} < x_2^{(m)} < \cdots < x_{n+1}^{(m)} = b$ such that

$$f(x_k^{(m)}) - f(x_{k-1}^{(m)}) \leq \frac{f(b) - f(a)}{n}.$$

Introduce

$$f^{(m)}(x) := f(x_k^{(m)}) + \frac{f(x_{k+1}^{(m)}) - f(x_k^{(m)})}{x_{k+1}^{(m)} - x_k^{(m)}}(x - x_k^{(m)}) \quad \text{if } x \in [x_k^{(m)}, x_{k+1}^{(m)}].$$

For any $x \in [x_k^{(m)}, x_{k+1}^{(m)}]$, it follows from the convexity of f that

$$\begin{aligned}|f^{(m)}(x) - f(x)| &= \Big|f(x_k^{(m)}) - f(x) + \frac{\big(f(x_{k+1}^{(m)}) - f(x_k^{(m)})\big)(x - x_k^{(m)})}{x_{k+1}^{(m)} - x_k^{(m)}}\Big| \\ &= f(x_k^{(m)}) - f(x) + \frac{\big(f(x_{k+1}^{(m)}) - f(x_k^{(m)})\big)(x - x_k^{(m)})}{x_{k+1}^{(m)} - x_k^{(m)}} \\ &\le \frac{\big(f(x_{k+1}^{(m)}) - f(x_k^{(m)})\big)(x - x_k^{(m)})}{x_{k+1}^{(m)} - x_k^{(m)}} \le \frac{f(b) - f(a)}{n},\end{aligned}$$

from which we know $\lim_{n\to\infty} |f^{(m)}(x) - f(x)| = 0$ for all $x \in [a, b]$. Lemma F.3 shows that $f^{(m)} \in H_{[a,b]}$ for all $m \in \mathbb{N}$. Therefore, f belongs to the closure of $H_{[a,b]}$. The proof is completed. □

Proof of Theorem F.1. According to the boundedness assumption of $f \in \mathcal{F}$ and the fact $\mathbf{0} \in \mathcal{F}$, there exist $B > 0$ such that

$$\begin{aligned}0 \le \min\Big\{ \sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i)), G\sqrt{2} \sup_{f\in\mathcal{F}} \sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\Big\} \\ \le \max\Big\{ \sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i)), G\sqrt{2} \sup_{f\in\mathcal{F}} \sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\Big\} \le B\end{aligned}$$

for all $\boldsymbol{\epsilon} \in \{\pm 1\}^n$. Let $t \in \mathbb{R}$ be an arbitrary number. Define $g_t : \mathcal{F} \mapsto \mathbb{R}$ by $g_t(f) = 0$ for any $f \neq \mathbf{0}$ and $g_t(\mathbf{0}) = t$. It is clear that

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \sup_{f\in\mathcal{F}} \big[g_t(f) + \sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i))\big] = \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \max\Big\{ \sup_{f\in\mathcal{F}: f\neq\mathbf{0}} \big[\sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i))\big], t\Big\}$$

and

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \sup_{f\in\mathcal{F}} \big[g_t(f) + G\sqrt{2}\sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\big] = \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \max\Big\{G\sqrt{2} \sup_{f\in\mathcal{F}: f\neq\mathbf{0}} \big[\sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\big], t\Big\}.$$

Plugging the above identities into (F.2) with $g = g_t$ gives

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \max\Big\{ \sup_{f\in\mathcal{F}: f\neq\mathbf{0}} \big[\sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i))\big], t\Big\} \le \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \max\Big\{G\sqrt{2} \sup_{f\in\mathcal{F}: f\neq\mathbf{0}} \big[\sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\big], t\Big\}.$$

If $t \ge 0$, the above inequality is equivalent to

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \max\Big\{ \sup_{f\in\mathcal{F}} \big[\sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i))\big], t\Big\} \le \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \max\Big\{G\sqrt{2} \sup_{f\in\mathcal{F}} \big[\sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\big], t\Big\} \tag{F.5}$$

by noting $\tilde{g}_i(\mathbf{0}) = 0$ for all $i \in \mathbb{N}_n$. If $t < 0$, it follows from (F.2) with $g(f) = 0$ that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \max\Big\{ \sup_{f\in\mathcal{F}} \big[\sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i))\big], t\Big\} \\ &= \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \sup_{f\in\mathcal{F}} \big[\sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i))\big] \le \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \sup_{f\in\mathcal{F}} \Big[G\sqrt{2}\sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\Big] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \max\Big\{G\sqrt{2} \sup_{f\in\mathcal{F}} \big[\sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\big], t\Big\},\end{aligned}$$

where we have used $\tilde{g}_i(\mathbf{0}) = 0$ for all $i \in \mathbb{N}_n$ in the first identity. That is, (F.5) holds for all $t \in \mathbb{R}$. Subtracting t from both sides of Eq. (F.5) gives

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n} \Big(\sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i \tilde{g}_i(f(\mathbf{x}_i)) - t\Big)_+ \le \mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}} \Big(G\sqrt{2} \sup_{f\in\mathcal{F}} \big[\sum_{i=1}^n \sum_{j=1}^d \epsilon_{i,j} f_j(\mathbf{x}_i)\big] - t\Big)_+, \quad \forall t \in \mathbb{R}, \tag{F.6}$$

from which we know

$$\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^n}\tilde{\tau}\Big(\sup_{f\in\mathcal{F}}\sum_{i=1}^n\epsilon_i\tilde{g}_i(f(\mathbf{x}_i))\Big)\leq\mathbb{E}_{\boldsymbol{\epsilon}\sim\{\pm1\}^{nd}}\tilde{\tau}\Big(G\sqrt{2}\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sum_{j=1}^d\epsilon_{i,j}f_j(\mathbf{x}_i)\Big),\quad\forall\tilde{\tau}\in H_{[0,B]}.$$

According to Lemma F.4, we know $\tau:[0,B]\to\mathbb{R}_+$ belongs to the closure of $H_{[0,B]}$. Therefore, Eq. (F.1) holds. The proof is completed. □

G. Lipschitz Continuity of Loss Functions

The following proposition is known in the literature (Lei et al., 2019). We prove it for completeness.

Proposition G.1. *(a) Let ℓ be defined as Eq. (3.2). Then ℓ is 1-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$ and 1-Lipschitz continuous w.r.t. $\|\cdot\|_2$.*

(b) Let ℓ be defined as Eq. (3.3). Then ℓ is 1-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$ and 1-Lipschitz continuous w.r.t. $\|\cdot\|_2$.

Proof. We first prove Part (a). For any $\mathbf{v}$ and $\mathbf{v}'$, we have

$$\begin{aligned}|\ell(\mathbf{v})-\ell(\mathbf{v}')| &= \big|\max\big\{0,1+\max_{i\in[k]}\{-v_i\}\big\}-\max\big\{0,1+\max_{i\in[k]}\{-v_i'\}\big\}\big|\\ &\leq|\max_{i\in[k]}\{-v_i\}-\max_{i\in[k]}\{-v_i'\}|\leq\max_{i\in[k]}|v_i-v_i'|=\|\mathbf{v}-\mathbf{v}'\|_\infty,\end{aligned}$$

where we have used the elementary inequality

$$|\max_{i\in[k]}a_i-\max_{i\in[k]}b_i|\leq\max_{i\in[k]}|a_i-b_i|.$$

This proves Part (a).

We now prove Part (b). It is clear that

$$\frac{\partial\ell(\mathbf{v})}{\partial v_i}=\frac{-\exp(-v_i)}{1+\sum_{i\in[k]}\exp(-v_i)}.$$

Therefore, the ℓ_1 norm of the gradient can be bounded as follows

$$\|\nabla\ell(\mathbf{v})\|_1\leq\frac{1}{1+\sum_{i\in[k]}\exp(-v_i)}\sum_{i\in[k]}\exp(-v_i)\leq1.$$

This proves Part (b). The proof is completed. □